RM-5888-PR
June 1969

THE DELPHI METHOD: AN EXPERIMENTAL STUDY OF GROUP OPINION

Norman C. Dalkey

prepared for
UNITED STATES AIR FORCE PROJECT RAND

Rand
SANTA MONICA, CA. 90406

Rand maintains a number of special, subject bibliographies containing abstracts of Rand publications in fields of wide current interest. The following bibliographies are available upon request:

Africa • Arms Control • Civil Defense • Combinatorics
Communication Satellites • Communication Systems • Communist China
Computing Technology • Decisionmaking • East-West Trade
Education • Foreign Aid • Health-related Research • Latin America
Linguistics • Long-range Forecasting • Maintenance
Mathematical Modeling of Physiological Processes • Middle East
Policy Sciences • Pollution • Procurement and R&D Strategy
Program Budgeting • SIMSCRIPT and Its Applications • Southeast Asia
Systems Analysis • Television • Urban Problems • USSR
Water Resources • Weather Forecasting and Control

To obtain copies of these bibliographies, and to receive information on how to obtain copies of individual publications, write to: Communications Department, Rand, 1700 Main Street, Santa Monica, California 90406.

Published by The Rand Corporation

RM-5888-PR
June 1969

THE DELPHI METHOD: AN EXPERIMENTAL STUDY OF GROUP OPINION

Norman C. Dalkey

This research is supported by the United States Air Force under Project Rand—Contract No. F44620-67-C-0045—Monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this study should not be interpreted as representing the official opinion or policy of Rand or of the United States Air Force.

Rand
SANTA MONICA, CA. 90406

PREFACE

This report deals with one aspect of RAND's continuing study of methods for improving decisionmaking. It describes the results of an extensive set of experiments conducted at RAND during the spring and summer of 1968. The experiments were concerned with evaluating the effectiveness of the Delphi procedures for formulating group judgments.

The study is of direct relevance for the use of experts as advisors in decisionmaking, especially in areas of broad or long-range policy formulation. For the Air Force, the results bear on methods of dealing with a wide spectrum of problems, ranging from long-term threat assessment to forecasts of technological and social development. Pilot studies of social and technological forecasts have been conducted by OAR and AFSC. The results presented in this Memorandum should increase the effectiveness of such studies in the future.

In industry, related techniques are being applied to both technological forecasting and the evaluation of corporate planning. Various public agencies are utilizing Delphi procedures for planning exercises related to education, health, and urban growth.

SUMMARY

The Delphi technique is a method of eliciting and refining group judgments. The rationale for the procedures is primarily the age-old adage "Two heads are better than one," when the issue is one where exact knowledge is not available. The procedures have three features: (1) Anonymous response—opinions of members of the group are obtained by formal questionnaire, (2) Iteration and controlled feedback—interaction is effected by a systematic exercise conducted in several iterations, with carefully controlled feedback between rounds. (3) Statistical group response—the group opinion is defined as an appropriate aggregate of individual opinions on the final round. These features are designed to minimize the biasing effects of dominant individuals, of irrelevant communications, and of group pressure toward conformity.

In the spring of 1968, a series of experiments were initiated at RAND to evaluate the procedures. The experiments were also designed to explore the nature of the information processes occurring in the Delphi interaction. The experiments were conducted using upper-class and graduate students from UCLA as subjects, and general information of the almanac type as subject matter. Ten experiments, involving 14 groups ranging in size from 11 to 30 members, were conducted. About 13,000 answers to some 350 questions were obtained.

The two basic issues being examined were (1) a comparison of face-to-face discussion with the controlled-feedback interaction, and (2) a thorough evaluation of controlled feedback as a technique of improving group estimates. The results indicated that, more often than not, face-to-face discussion tended to make the group estimates less accurate, whereas, more often than not, the anonymous controlled feedback procedure made the group estimates more accurate. The experiments thus put the application of Delphi techniques in areas of partial information on much firmer ground.

Of greater long-range significance is the insight gained into the nature of the group information processes. Delphi procedures create a well-defined process that can be described quantitatively. In particular, the average error on round one is a linear function of the dispersion of the answers. The average amount of change of opinion between round one and round two is a well-behaved function of two parameters—the distance of the first-round answer from the group median, and the distance from the true answer.

Another result of major significance is that a meaningful estimate of the accuracy of a group response to a given question can be obtained by combining individual self-ratings of competence on that question into a group rating. This result, when combined with the relationship between accuracy and standard deviation mentioned above, opens the possibility of attaching accuracy scores to the products of a Delphi exercise.

A number of supplementary analyses—including the effect of time-to-answer on accuracy, the comparison of performance as a function of college major, and the effect of different question format—have added useful elements to the overall picture, giving additional weight to the presumption that information-handling procedures that are appropriate for well-confirmed material are not suitable for the less well confirmed area of expert opinion.

Although the experiments conducted to date have been informative beyond initial expectations, they represent only a small beginning in a field of research that could be called "opinion technology."

CONTENTS

1. THE SPECTRUM OF DECISION INPUTS

One of the thorniest problems facing the policy analyst is posed by the situation where, for a significant segment of his study, there is unsatisfactory information. The deficiency can be with respect to data—incomplete or faulty—or more seriously with respect to the model or theory—again either incomplete or insufficiently verified. This situation is probably the norm rather than a rare occurrence.

The usual way of handling this problem is by what could be called "deferred consideration." That is, the analyst carries out his study using whatever good data and confirmed models he has and leaves the "intangibles" to the step called "interpretation of results."* In some cases the deferment is more drastic. The analyst presents his study, for what it is worth, to a decisionmaker, who is expected to conduct the interpretation and "inclusion in the total picture."

In describing the interpretation-of-results step, interesting words are likely to appear. These include terms like "judgment," "insight," "experience," and especially as applied to decision-makers, "wisdom" or "broad understanding." These terms contrast with the presumed precision, scientific care, and dependence on data that characterize operations research. Above all, there is a slightly mystical quality about the notions. They are never explained. Standards of excellence are lacking. And there is more than a hint that the capabilities involved somehow go beyond the more mundane procedures of analysis.

---

*The not infrequent case where the analyst "makes do" with faulty data or shaky models has been sufficiently excoriated in the manuals of operations research methodology.

Taking a look at the kinds of information that can play a role in decisionmaking, there are roughly three types (see Fig. 1). On the one hand, there are assertions that are highly confirmed—assertions for which there is a great deal of evidence backing them up. This kind of information can be called knowledge. At the other end of the scale is material that has little or no evidential backing. Such material is usually called speculation. In between is a broad area of material for which there is some basis for belief but that is not sufficiently confirmed to warrant being called knowledge. There is no good name for this middling area. I call it opinion. The dividing lines between these three are very fuzzy, and the gross trichotomy smears over the large differences that exist within types. However, the three-way split has many advantages over the more common tendency to dismiss whatever is not knowledge as mere speculation.

Where in this scale do the products of judgment, wisdom, insight, and similar intellectual processes, lie? Not in speculation, we hope. And, almost by definition, not in knowledge. The most reasonable interpretation would be that these are flattering names for kinds of opinion.*

Unfortunately, there is no practical, objective measure for the dimension of evidence sketched in Fig. 1. The best we have is an intuitive and rough feeling for the scale.** The prototype of knowledge may be found in the systematized, experimentally confirmed propositions of the natural sciences. But many of the assertions in the area that is called "common sense" have an equal solidity; e.g., the gross features of

---

*One might say, "Wisdom is opinion with charisma."

**A Delphi approach for locating assertions on the evidence scale will be discussed in Section 9.8, p. 68ff.

SPECTRUM OF INPUTS

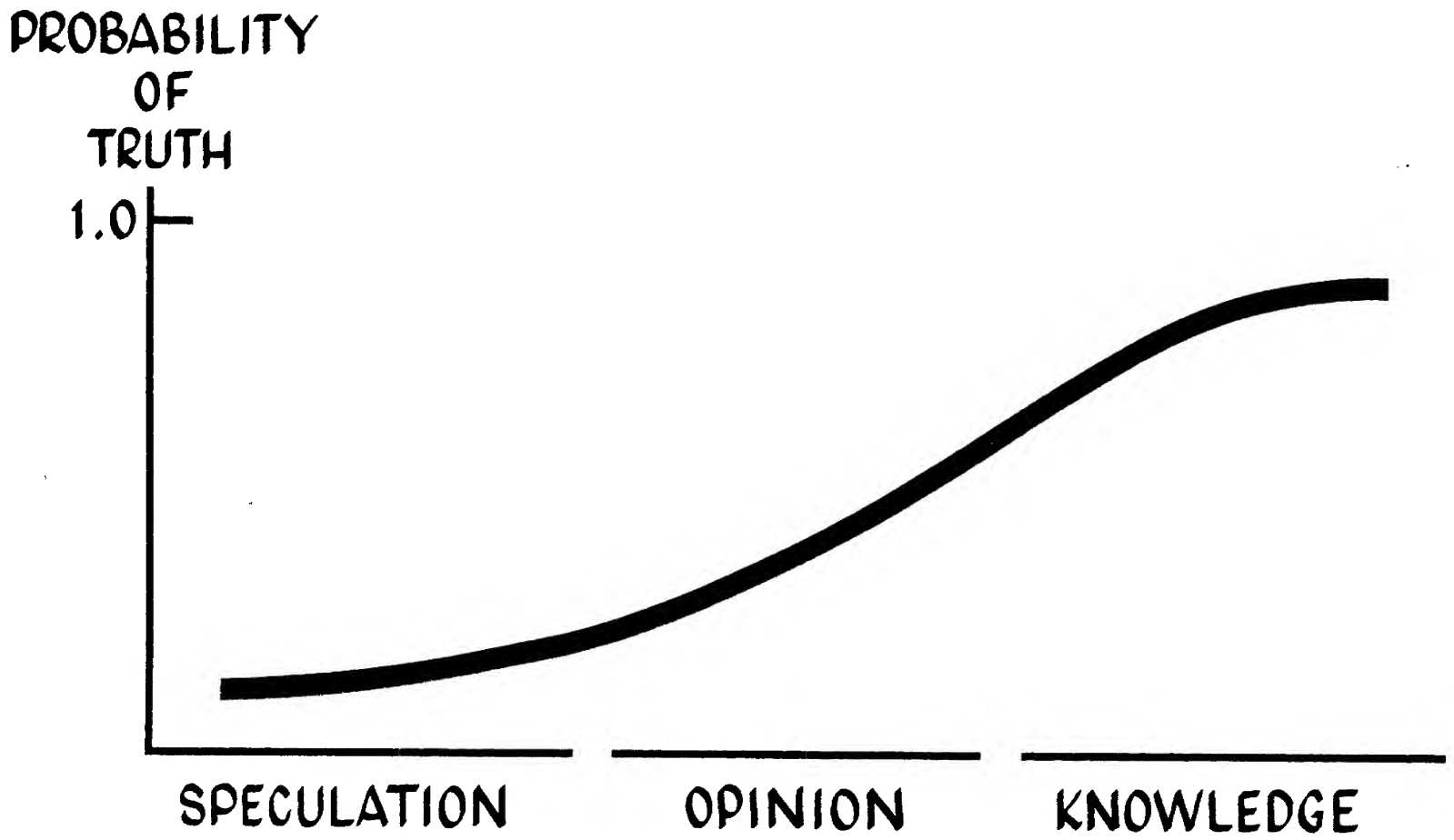

Fig. 1

gravity—"unsupported objects fall to the surface of the earth", the permanence of objects, and the like. A large part of the empirical generalizations of common technology are equally well confirmed. The technologist's criterion—does it work?—is at least as effective in eliminating unfounded notions as the scientist's is it confirmed by laboratory experiment?

In the following it will be taken for granted that methods of dealing with material in the area of knowledge are in reasonably good order. There are, of course, many problems of detail—the warrantability of extrapolation, the application of statistical measures where underlying distributions are unknown, and the like. But these difficulties are small compared with the conceptual vacuum that appears to exist in the area of opinion.

With respect to speculation, it appears very difficult to say anything wise other than to avoid it whenever possible. That isn't very helpful. It appears likely that most major policy decisions involve more than a dash of speculative inputs. Some of the general results described below are applicable to speculation, but how useful it is to the decisionmaker to furnish him with refined speculation is hard to say.

This report sidesteps the even more difficult issue raised by the fact that most practical decision situations involve a mixture of all three types of information. The delicate balancing of the weight to give each kind of material is a second-level sort of "wisdom" that has not yet been investigated.

In discussions of policy analysis it is usual to distinguish two kinds of assertions, factual statements and value judgments. It is an open question whether there is any basic conceptual difference between these two, but there are certainly very large practical differences.

In particular, value judgments tend to be much vaguer and displaced toward the opinion and speculative and of the solidity scale. The experimental results described below are concerned with factual material, but there is a short comment on value judgments in Section 10.

With respect to factual statements, it is worth pointing out that the crude scale of "solidity" is related to the likelihood that assertions are true. In the area of knowledge, by definition the probability of an assertion being true is relatively high; for speculative material the probability is low; and for opinion it is middling (see Fig. 1). This point is rather vital. There is an irrepressible urge on the part of analysts to move the arena of action entirely into the knowledge area. Sometimes this is possible. In general, it is not. When an opinion is expressed, it is an inescapable fact of life that whatever is said, there is a reasonable probability of its being false.

2. TWO HEADS ARE BETTER THAN ONE

There is a kind of technology for dealing with opinion that has been applied throughout historical times and probably in more ancient times as well. The technology is based on the adage "Two heads are better than one," or more generally "n heads are better than one." Committees, councils, panels, commissions, juries, boards, the voting public, legislatures ....the list is long, and illustrates the extent to which the device of pooling many minds has permeated society.*

The basis for the n-heads rule is not difficult to find. It is a tautology that, on any given question, there is at least as much relevant information in n heads as there is in any one of them. On the other hand, it is equally a tautology that there is at least as much misinformation in n heads as there is in one. And it is certainly not a tautology that there exists a technique of extracting the information in n heads and putting it together to form a more reliable opinion. With a given procedure, it may be the misinformation that is being aggregated into a less reliable opinion.

The n-heads rule, then, depends upon the procedures whereby the n heads are used. There is one kind of procedure and one kind of factual judgment where the n-heads rule comes very close to a tautology. Consider the case where the judgment required is a numerical estimate—e.g.,

---

*Most of these groups have more than one function. They can operate to transmit information, to coordinate action, to diffuse responsibility, to formulate policy, etc. All of these functions are important. None of the discussion below should be taken to apply directly to these other functions. In the present context we are concerned with the use of groups to formulate factual judgments. If the results of the present study appear suggestive with regard to the other functions of groups, I can only hope that this tends to generate additional experimentation.

the date at which a certain technological development will occur, or the size of world population in 1990—and assume you have a group of indistinguishable experts with respect to this estimate; that is, you have no way of asserting that one expert is more knowledgeable than another. Is it better to select the opinion of one expert at random or to take some statistical aggregate of the opinions of the group? It is a near-tautology that you are at least as well off to take the mean or the median as to select an expert at random.* This is, of course, a very weak statement. It can be most simply illustrated by using the median as the statistical representative of the group answer. Referring to Fig. 2, it is clear than, independent of the distribution of answers, and independent of the location of the true answer T, the median of the individual answers M is at least as close to the true answer as <u>one-half of the group</u>. If the range of group answers includes the true, then, in general, the median is closer to the true answer than more than half of the group, as in Fig. 3.

In practical situations, the range of answers is very likely to include the true answer, in which case the stronger assertion is valid. Fig. 4 shows the dependence on group size of the mean accuracy of a group response for a large set of experimentally derived answers to factual questions. The curve was derived by computing the average error of groups of various sizes where the individual answers were drawn from the experimental distribution. The error is

---

*The precise statement is: for the median, the probability that the median is at least as close to the true answer as any individual response is at least one half; for the mean, the error of the mean, (measured by the distance to the true answer) is less than or equal to the average error of the individual answers. These two criteria are not equivalent, and for different decision situations one or the other could be more appropriate.

WORST CASE : MEDIAN BETTER THAN HALF

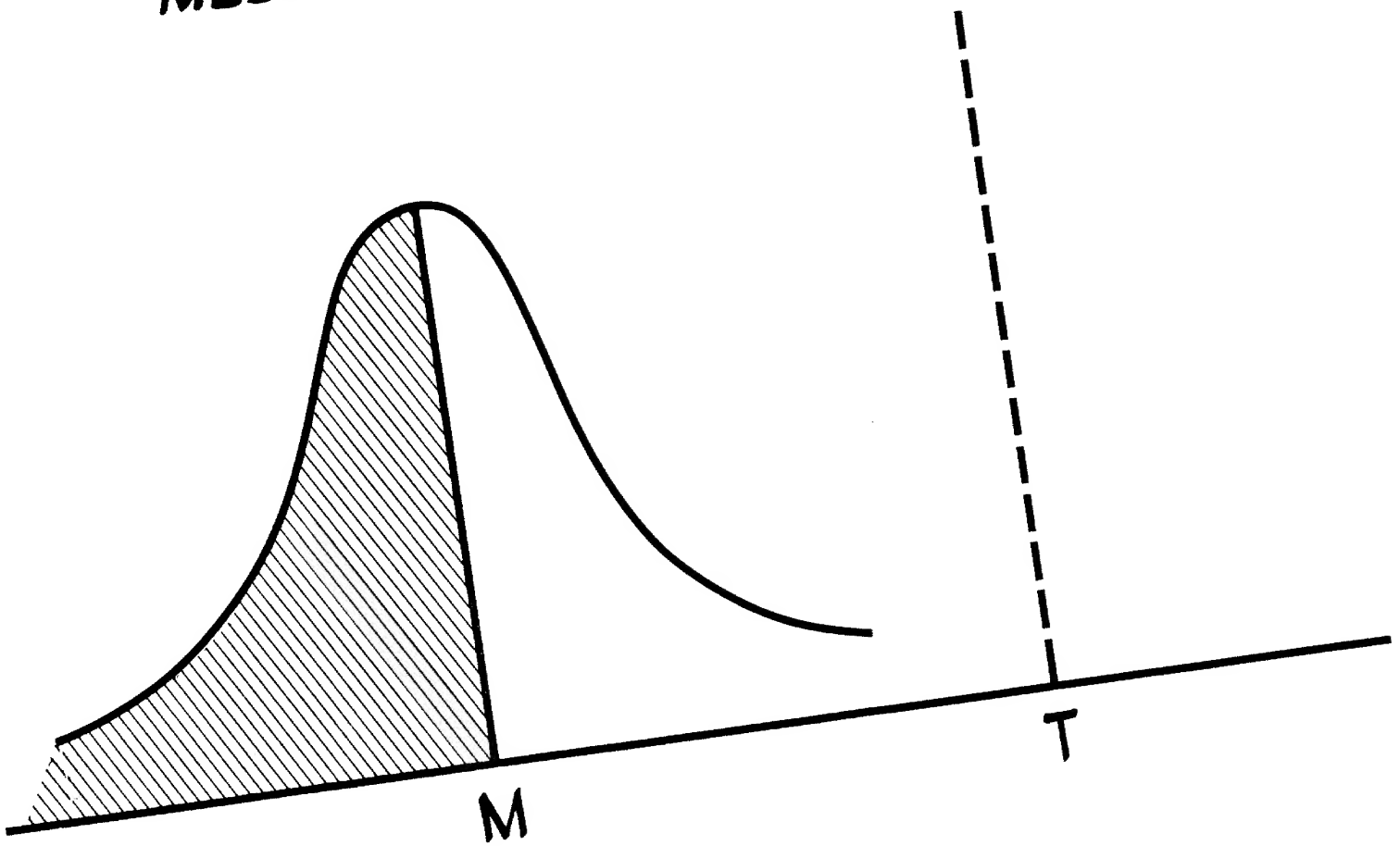

Fig. 2

NORMAL CASE:
MEDIAN BETTER THAN MORE THAN HALF

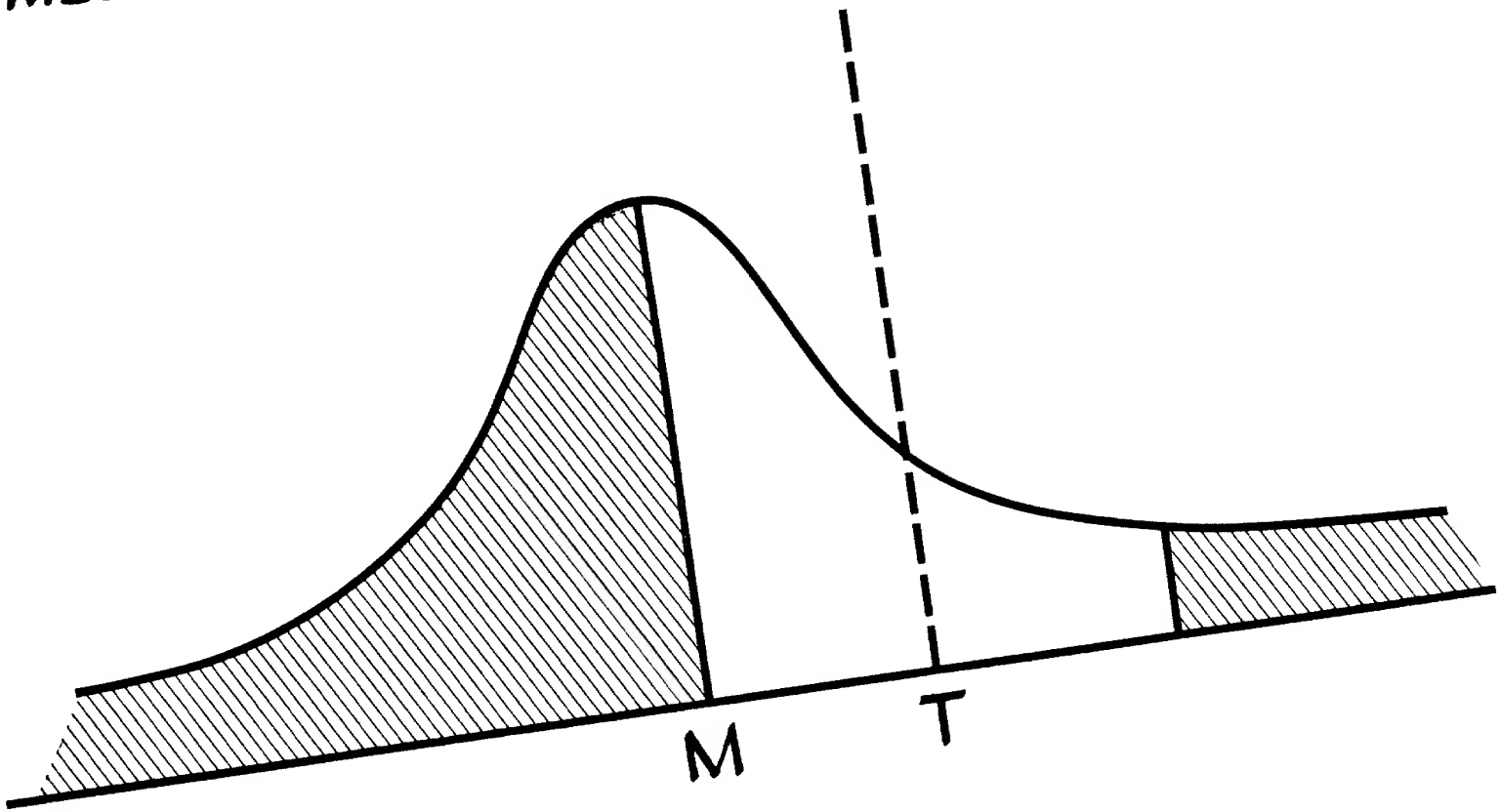

Fig. 3

measured on a logarithmic scale.* It is clear from Fig. 4 that with this population of answers, the gains in increasing group size are quite large. It is interesting that the curve appears to be decreasing in a definite fashion, even with groups as large as 29. This was the largest group size we used in our experiments.

Another important consideration with respect to the n-heads rule has to do with reliability. The most uncomfortable aspect of opinion from the standpoint of the decisionmaker is that experts with apparently equivalent credentials (equal degrees of expertness) are likely to give quite different answers to the same question. One of the major advantages of using a group response is that this diversity is replaced by a single representative opinion.** However, this feature is not particularly interesting if different groups of experts, each made up of equally competent members, come up with highly different answers to the same question.

In general, one would expect that in the area of opinion group responses would be more reliable than individual opinions, in the simple sense that two groups (of equally competent experts) would be more likely to evidence similar answers to a set of related questions than would two

---

*These were questions where the experimenters knew the answer but the subjects did not. The group error is the absolute value of the natural logarithm of the group median divided by the true answer. The groups used to construct Fig. 4 were "synthetic"; i.e., they were randomly selected sets of answers of the appropriate number drawn from the experimental distributions of answers.

**Whether this is the best use of group opinion, or whether the decisionmaker should take into account the full distribution of answers, and also make use of ranges of uncertainty on the part of individual respondents is an important topic in its own right, that will be partially explored in later sections.

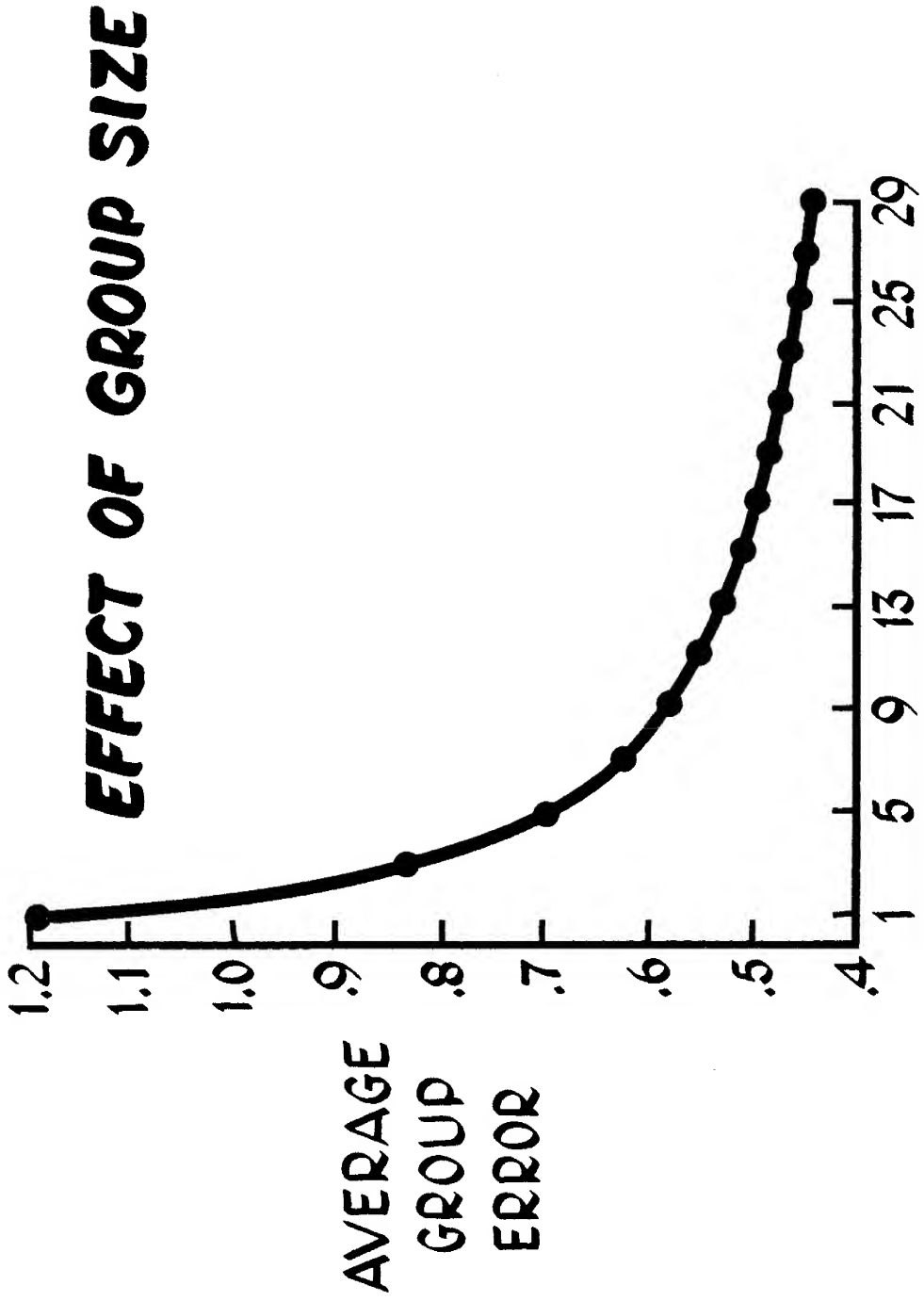
EFFECT OF GROUP SIZE
AVERAGE GROUP ERROR
1.2
1.1
1.0
.9
.8
.7
.6
.5
.4
1
5
9
13
17
21
25
29
NUMBER IN GROUP

Fig. 4

individuals. This "similarity" can be measured by the correlation between the answers of the two groups over a set of questions. But the assertion that groups will be more reliable than individuals is not a tautology. It depends on the distributions of answers that would be obtained from the total population of potential respondents, and it depends upon the method of selecting the subgroups out of this population. The result can be expected to hold if the distributions of answers for the potential population are not highly distorted, and if the subgroups are selected at random. There are clearly implications of this remark for the rules for selecting members of advisory bodies—in practice small advisory groups are probably never selected at random out of the total potential pool of experts.

For the analyst using expert opinion within a study, reliability can be considered to play somewhat the same role as reproducibility in experimental investigations. It is clearly desirable for a study that another analyst using the same approach (and different experts) arrive at similar results.

Fig. 5 shows the relationship between reliability and group size for the experimental population of answers to questions already mentioned. It was constructed by selecting at random pairs of groups of respondents of various sizes and correlating the median responses of the pairs on twenty questions. The ordinate is the average of these correlations.

It is clear that there is a definite and monotonic increase in the reliability of the group responses with increasing group size. It is not clear why the relationship would be approximately linear between n = 3 and n = 11.

In the area of opinion, then, the n-heads rule appears to be justified by considerations of both improved average accuracy, and reliability. The question remains whether

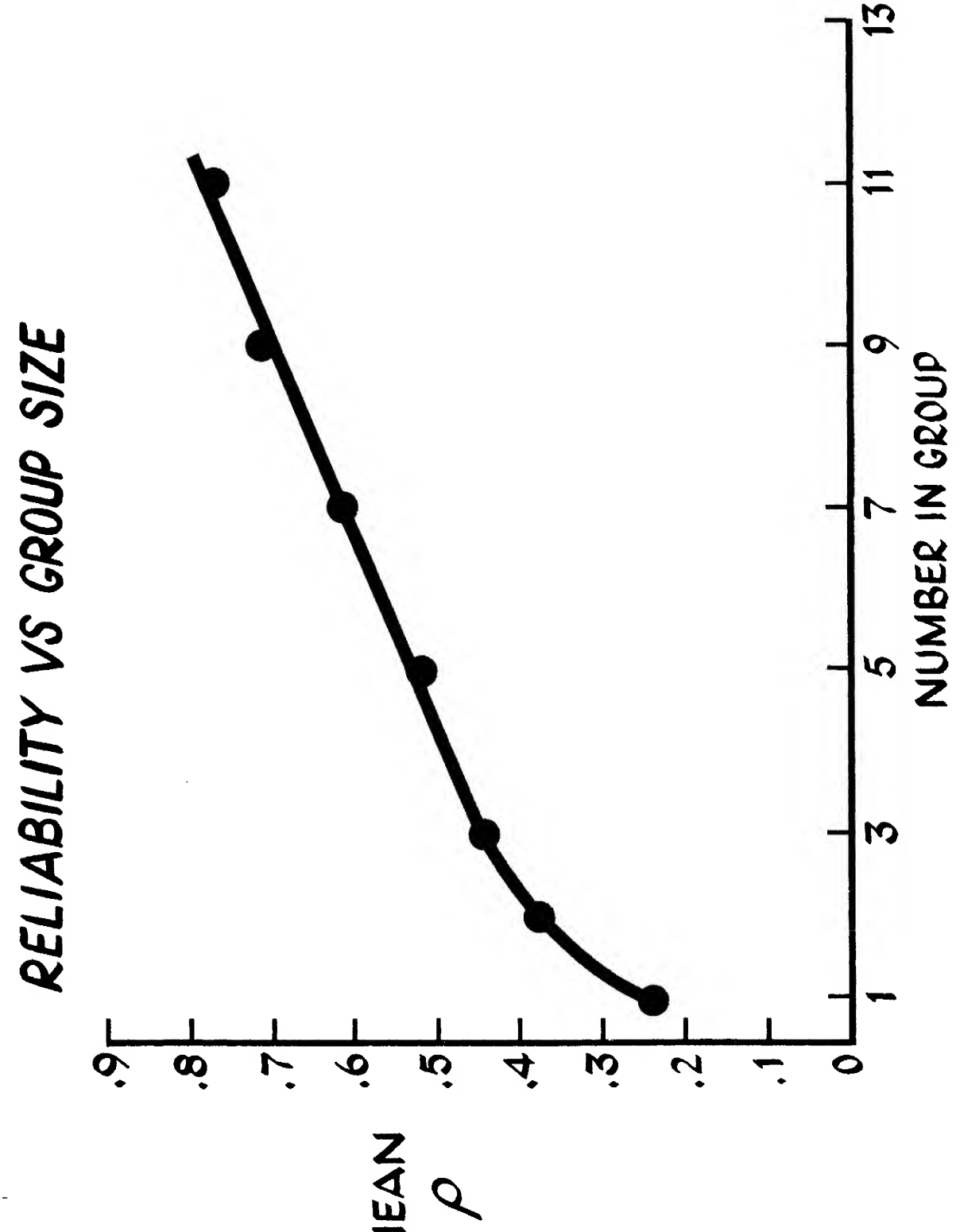
RELIABILITY VS GROUP SIZE
MEAN
ρ
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
1
3
5
7
9
11
13
NUMBER IN GROUP

Fig. 5

these quasi-statistical properties of group opinion can be improved upon by allowing more direct pooling of information on the part of the group.

The traditional way of pooling individual opinions is by face-to-face discussion. Numerous studies by psychologists in the past two decades have demonstrated some serious difficulties with face-to-face interaction [2]. Among the most serious are: (1) Influence of dominant individuals. The group opinion is highly influenced, for example, by the person who talks the most. There is very little correlation between pressure of speech and knowledge. (2) Noise. By noise is not meant auditory level (although in some face-to-face situations this may be serious enough!) but semantic noise. Much of the "communication" is a discussion group has to do with individual and group interests, not with problem solving. This kind of communication, although it may appear problem oriented, is often irrelevant or biasing. (3) Group pressure for conformity. The experiments of Asch [3] demonstrate in dramatic fashion the distortions of individual judgment that can occur from group pressure.

In experiments at RAND and elsewhere, it has turned out that, after face-to-face discussion, more often than not the group response is less accurate than a simple median of individual estimates without discussion.

3. DELPHI

There has been a somewhat intermittent series of studies at The RAND Corporation since its early days concerned with the problem of using group information more effectively. The early studies were concerned mainly with improving the statistical treatment of individual opinions [4]. They indicated that some formal properties of individual estimates (precision, definiteness) could be used to rate the success of short-term predictions, and that background information (as measured by a standard achievement test) had a small but significant influence on the success of predictions. Both of these effects were fairly well washed out by combining estimates into group predictions.

In 1953, Dalkey and Helmer [5] introduced an additional feature, namely iteration with controlled feedback. The set of procedures that have evolved from this work has received the name "Delphi"—a somewhat misleading appelation, since there is little that is oracular about the methods.

The Delphi procedures received a very large boost in general interest with the publication of Gordon and Helmer's study of forecasting technological events [6]. In the area of long-range forecasting, it is difficult to dodge the fact that a large part of the activity is at least within the area of opinion, and possibly worse. That particular study happened to coincide with a surge of interest in long-range forecasting itself, with an attendant interest in the systematic use of expert opinion.

In the last three years there has been a very large increase in applications of the procedures, primarily by industry for the forecasting of technological developments [7], but also by a variety of organizations for exploring policy decisions in areas such as education, public transportation, public health, etc. At present, it is difficult to obtain a clear picture of how widespread the applications are; but a crude guess would put the number of studies recently completed, under way, or in the planning stages at well over a hundred.

In light of this widespread exploitation, the question of just how effective the procedures are has considerable practical import.

In general, the Delphi procedures have three features: (1) anonymity, (2) controlled feedback, and (3) statistical group response. Anonymity, effected by the use of questionnaires or other formal communication channels, such as on-line computer communication, is a way of reducing the effect of dominant individuals. Controlled feedback—conducting the exercise in a sequence of rounds between which a summary of the results of the previous round are communicated to the participants—is a device for reducing noise. Use of a statistical definition of the group response is a way of reducing group pressure for conformity; at the end of the exercise there may still be a significant spread in individual opinions. Probably more important, the statistical group response is a device to assure that the opinion of every member of the group is represented in the final response. Within these three basic features, it is, of course, possible to have many variations.

There are several properties of a Delphi exercise that should be pointed out. The procedure is, above all, a rapid and relatively efficient way to "cream the tops of the heads" of a group of knowledgeable people. In general,

it involves much less effort for a participant to respond to a well-designed questionnaire than, for example, to participate in a conference or to write a paper. A Delphi exercise, properly managed, can be a highly motivating environment for respondents. The feedback, if the group of experts involved is mutually self-respecting, can be novel and interesting to all. The use of systematic procedures lends an air of objectivity to the outcomes that may or may not be spurious, but which is at least reassuring. And finally, anonymity and group response allow a sharing of responsibility that is refreshing and that releases from the respondents inhibitions. I can state from my own experience, and also from the experience of many other practitioners, that the results of a Delphi exercise are subject to greater acceptance on the part of the group than are the consensuses arrived at by more direct forms of interaction.

I believe all of these features of a Delphi exercise are desirable, especially if the exercise is conducted in the context of policy formulation where group acceptance is an important consideration. Like any technique for group interaction, the Delphi procedures are open to various misuses; much depends on the standards of the individual or group conducting the exercises.

4. EXPERIMENTS

In addition to questioning the effects on free expression of opinion and group acceptance, it still must be asked whether the use of iteration and controlled feedback have anything to offer over the "mere" statistical aggregation of opinions. I put "mere" in quotation marks; in the area of opinion much can be gained by the simple arithmetical pooling of individual opinions as shown above. To get some measure of the value of the procedures, and also to obtain, as a basis for improving the procedures, some insight into the information processes that occur in a Delphi exericse, we undertook a rather extensive series of experiments at RAND starting in the spring of 1968.* We used upper-class and graduate students, primarily from UCLA, as subjects. They were paid for their participation. For subject matter we chose questions of general information, of the sort contained in an almanac or statistical abstract. Typical questions were: "How many telephones were in use in Africa in 1965?" "How many suicides were reported in the U.S. in 1967?" "How many women marines were there at the end of World War II?" This type of material was selected for a variety of reasons: (1) We wanted questions where the subjects did not know the answer but had sufficient background information so they could make an informed estimate. (2) We wanted questions where there was a verifiable answer to check the performance of individuals and groups. (3) We wanted questions with numerical answers to a reasonably wide range of performance could be scaled. As far as we

---

*The team involved in these experiments consisted, in addition to myself, of Bernice Brown, Tom Brown, Samuel Cochran, Olaf Helmer and Richard Rochberg. The fruitfulness of the experimental program is directly ascribable to the high level of competence of these co-workers.

can tell, the almanac type of question fits these criteria quite well. There is the question whether results obtained with this very restricted type of subject matter apply to other kinds of material. We can say that the general-information type of question used had many of the features ascribable to opinion: namely, the subjects did not know the answer, they did have other relevant information that enabled them to make estimates, and the route from "other relevant information" to an estimate was neither immediate nor direct.*

For about half of the experiments, the design called for a control group and an experimental group, each of about 15 subjects. For the others, the iterative structure allowed the group to be its own control. The experiments were conducted as closed information sessions; no inputs beyond the background information of the subjects were introduced. The standard task was answering 20 questions of an almanac sort. The questions were different from experiment to experiment (to preclude inadvertent transfer of information outside the experiments). The basic feedback between rounds was the median and the upper and lower quartiles of the previous-round answers. Additional feedback, summarized from subject responses, was introduced in some cases for experimental evaluation. Altogether, there were 11 experiments, involving close to 5000 answers to some 300 questions on each of several rounds. I will not describe

---

*The results from other experiments using as subject matter short-range prediction of economic, technological, and social events [8,4] appear to substantiate the assumption that there is very little difference between the general properties of answers to our estimation-type questions and the short-range predictions; e.g., with respect to distribution of answers, convergence on feedback, relative accuracy of individual and group responses, etc. However, this observation should be confirmed with more controlled exercises.

all the details of each experiment but will present a resume of the major results.*

The general outcome of the experiments can be summarized roughly as follows: (1) On the initial round, a wide spread of individual answers typically ensued. (2) With iteration and feedback, the distribution of individual responses progressively narrowed (convergence). (3) More often than not, the group response (defined as the median of the final individual responses) became more accurate. This last result, of course, is the most significant. Convergence would be less than desirable if it involved movement away from the correct answer.

---

*Details of procedure, the list of questions employed, and specific outcomes of the experiments are contained in [9].

5. COMPARISON OF FACE-TO-FACE AND ANONYMOUS INTERACTION

Two experiments were devoted to comparing the performance of groups using face-to-face discussion with groups employing anonymous, questionnaire-feedback interaction. The first experiment involved ten graduate student summer consultants to The RAND Corporation. These were divided into two groups of five, and the twenty questions were presented in four blocks of five each, following an ABBA design—A denoting face-to-face discussion and B denoting questionnaire feedback for one group, with the reverse for the other group. Thus, each group answered ten questions in discussion sessions and ten in questionnaire sessions. During questionnaire sessions, the subjects remained in separate cubicles. Approximately two and a half hours on successive afternoons were used to answer each block of five questions for each method of interaction.

The face-to-face groups were instructed to follow a specific procedure in dealing with each question. This procedure involved selection of a discussion leader at random, listing all information known to the group relevant to the question, devising several different ways of answering the question from the listed information, producing estimates by each of these ways, evaluating the relative solidity of each approach, and if possible, reaching a group consensus on the answer. For all but one of the twenty questions, a group consensus was arrived at.

The questionnaire procedure involved four rounds of estimates, feedback of medians and quartiles from the previous rounds, and reestimates. In addition, on some of

the rounds the subjects were asked to rate their competence on the questions and to submit reasons for their answers. These additional features will be discussed in a later section.

The basic result was that the median response of the questionnaire group was more accurate in 13 cases, and the consensus of the face-to-face group was more accurate in 7 cases. Considered as an isolated experiment, this result is not statistically significant, a fact that is borne out by an analysis of variance.* However, when this experiment is considered along with several others showing the same kind of outcome, the results appear more significant.

The second experiment used a different design. We considered the possibility that 5 subjects were already a "large" group for face-to-face discussion. Accordingly, we took a group of 23 respondents, obtained their initial estimates on 20 questions individually, and then divided them into 7 groups of 3 and one group of 2. The medians and quartiles of the total group on the first round were fed back and each subject again made an individual estimate. The small groups then discussed the questions one at a time and again made individual estimates for each question. In the instructions for the discussion groups, some of the difficulties with face-to-face interaction that had been identified in experiments at RAND and elsewhere were outlined, and the groups were requested to guard against the biasing influences whenever possible.

The basic outcome of this experiment is given in Table 1.

---

*An analysis of sums of ranks did reveal a statistically significant lower sum of ranks for the questionnaire group.

Table 1

COMPARISON OF ACCURACY OF GROUP MEDIANS
AFTER CONTROLLED FEEDBACK AND AFTER DISCUSSION

	Change Between Rounds 1 and 2 Delphi	Change Between Rounds 2 and 3 Discussion	Change Between Rounds 1 and 3
More Accurate	8	9	11
Same	8	3	0
Less Accurate	4	8	9

The picture presented by these results is not as clear cut as that from the first experiment. The improvement (difference between more and less accurate) between rounds one and two is somewhat greater than the improvement between rounds one and three. From this point of view, the overall improvement would have been greater without the discussion.

The outcomes of these two experiments are in accord with the results obtained by Campbell [7]. In Campbell's study, a group of graduate students, some with business experience, first underwent an exercise designed to compare Delphi procedures with "normal" procedures for making short-range forecasts of a set of 16 economic indices. The "normal" procedures were not defined—whatever methods the subjects wished to use. Free communication was allowed the non-Delphi groups. The Delphi forecasts were, on the final (fourth) round, more accurate in 13 cases; the normal procedures were more accurate in 2 cases. This result is highly favorable with respect to the comparison of systematic and controlled interaction as against informal interaction.

Of more direct relevance to the comparison of face-to-face and Delphi procedures was an exercise Campbell conducted at the end of the Delphi study. The teams were called together in face-to-face sessions and requested to

discuss either four or five (depending on the time available) of the sixteen indices and come up with a consensus answer. For the two groups that had engaged in the Delphi exercises, the post-discussion led to a degradation of answers in three out of four and in four out of five cases. On the other hand, the unstructured interaction groups profited by the discussion in three out of four and in three out of five cases.

Although these three experiments do not yield a clear and simple outcome, the negative conclusion that discussion does not display an advantage over statistical aggregation appears well confirmed; and the overall weight of the experiments tends to confirm the hypothesis that, more often than not, discussion leads to a degradation of group estimates. However, further experiments are desirable to establish the effect of face-to-face discussion more firmly.

6. THE NATURE OF ESTIMATION

One of our basic interests was to obtain a better understanding of the estimation process itself. The experiments were not designed specifically to explore this subject, but we hoped that the data would furnish some insights. As it turned out, the experimental data were very revealing.

The distribution of individual first round-answers for twelve of the experimental groups is displayed in Fig. 6. In constructing this chart, the responses were normalized so that the mean and standard deviation of the logarithms of the responses of the group to each question were zero and one, respectively. The distribution begins at zero, since all questions asked (except for an inadvertant one concerning the lowest temperature ever recorded in Florida) have non-negative answers. Drawn on the same chart is a log-normal curve with mean and standard deviation of one. The distribution of first-round answers is impressively log-normal.

The distribution warrants two comments. (1) The range of answers is rather astonishing. Although not directly readable from the chart, answers to the same question often differ by a factor of 10^4. (2) The log-normality of the distribution indicates that the subjects were thinking as much in terms of ratios, or so to speak, in terms of the size of the answer, as they were about the precise magnitude of the answer. This suggested that a reasonable scaling of individual answers was a logarithmic transformation. This scaling has been used in most of the analyses of the data.

Figure 7 displays the distribution of second-round answers, where a log-normal curve has been appended for comparison. The changes are manifest—there has been a large shift toward the mean, and the curve is distinctly

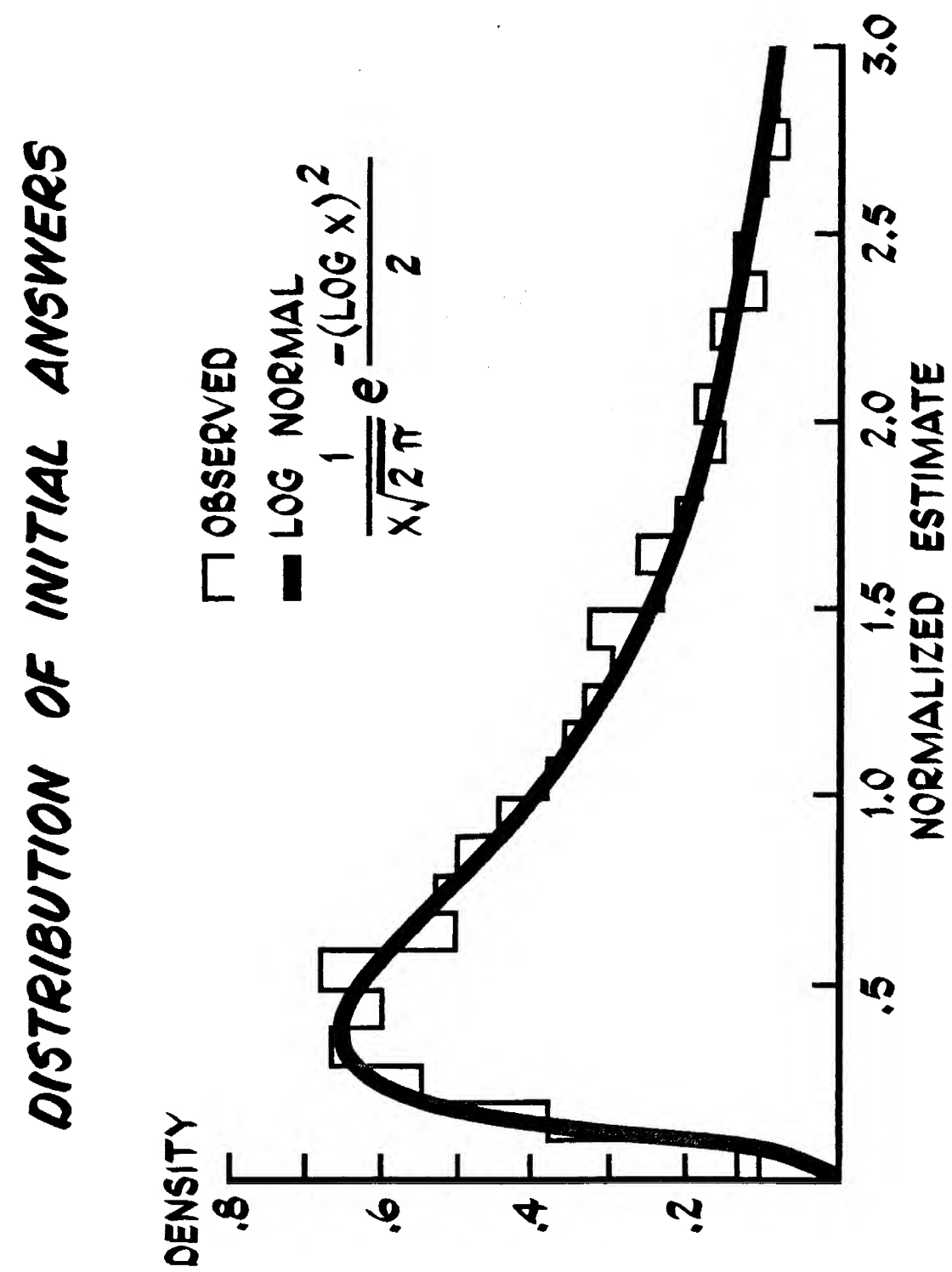
DISTRIBUTION OF INITIAL ANSWERS
OBSERVED
LOG NORMAL
$\frac{1}{x\sqrt{2\pi}} e^{\frac{-(\log x)^2}{2}}$
DENSITY
.8
.6
.4
.2
.5
1.0
1.5
2.0
2.5
3.0
NORMALIZED ESTIMATE

Fig. 6

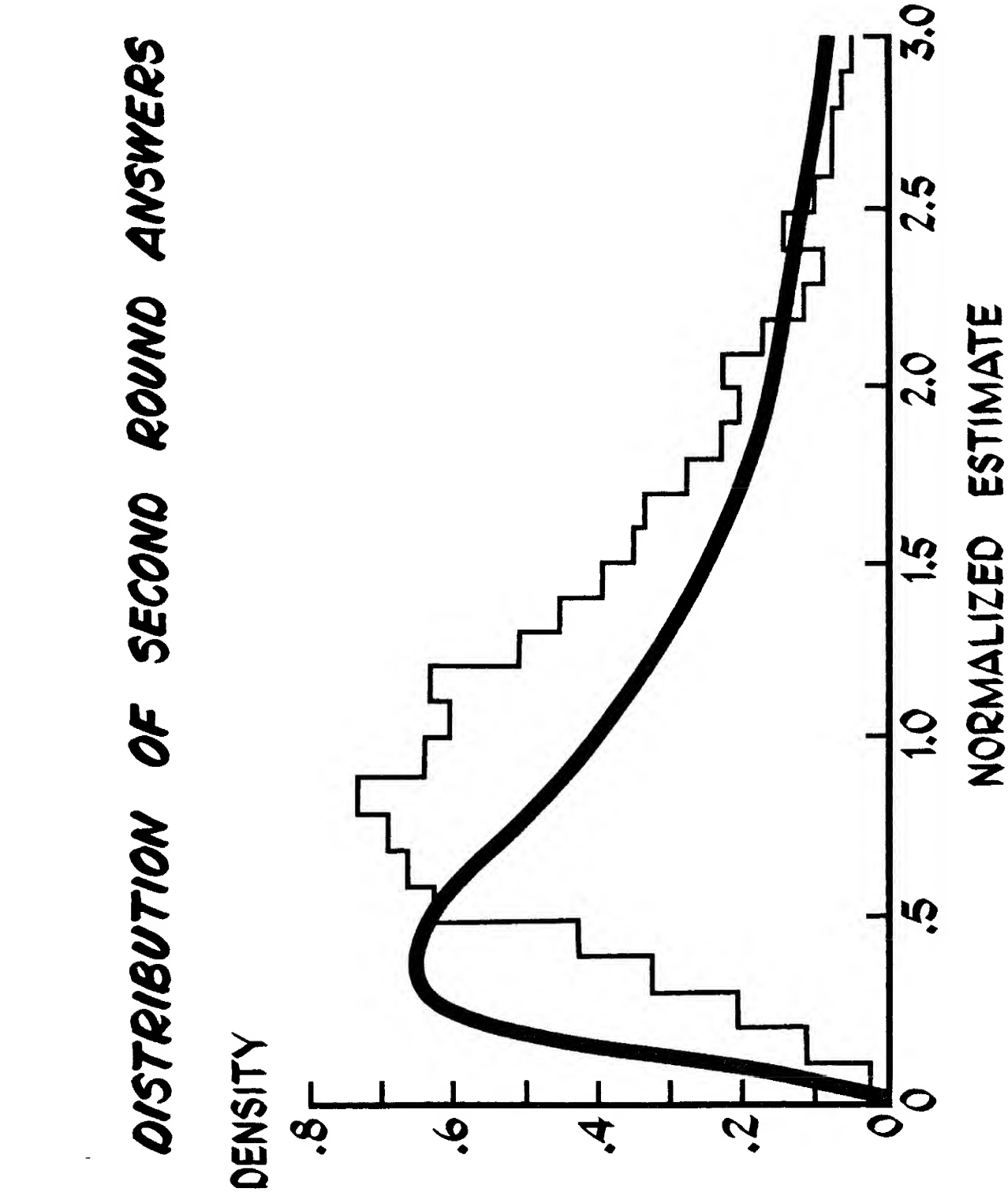
DISTRIBUTION OF SECOND ROUND ANSWERS
DENSITY
.8
.6
.4
.2
0
.5
1.0
1.5
2.0
2.5
3.0
NORMALIZED ESTIMATE

Fig. 7

different from the log-normal. The shift toward the mean represents a convergence of answers toward the group response. However, the convergence is by no means complete; the second-round distribution still has a large range.

It would be a step forward if we could assume that the "underlying" distribution of initial answers to each question were of a similar shape. The best we can say at present is that the observed distributions of answers on individual questions are compatible with the assumption that they are log-normal. It would be even more significant for further investigations if we could use the data to say something about the existence of, and the nature of, distributions "in the minds" of each respondent on individual questions. The data are, of course, compatible with the assumption that such distributions exist and are log-normal, but there is no direct way to verify the assumption from present information.

Figure 6 suggests that there is a measure of order underlying the superficially chaotic set of answers obtained on individual questions. A salient issue in this regard concerns the relation of the accuracy of responses and the amount of agreement within the group. A widespread, but intuitive, belief is that if a group displays a fair amount of agreement they are more likely to be correct than if they exhibit a wide spread of answers. Our initial attempts to test this hypothesis by computing the correlation between spread (measured by the standard deviation of answers on a given question) and accuracy (measured by the absolute value of the logarithm of the group median divided by the true answer) produced a disappointing result. The correlation turned out to be .26—statistically significant, but not high enough to be interesting.

However, when a plot is made of the average error as a function of standard deviation, we obtain the set of points displayed in Fig. 8, approximated by the upper

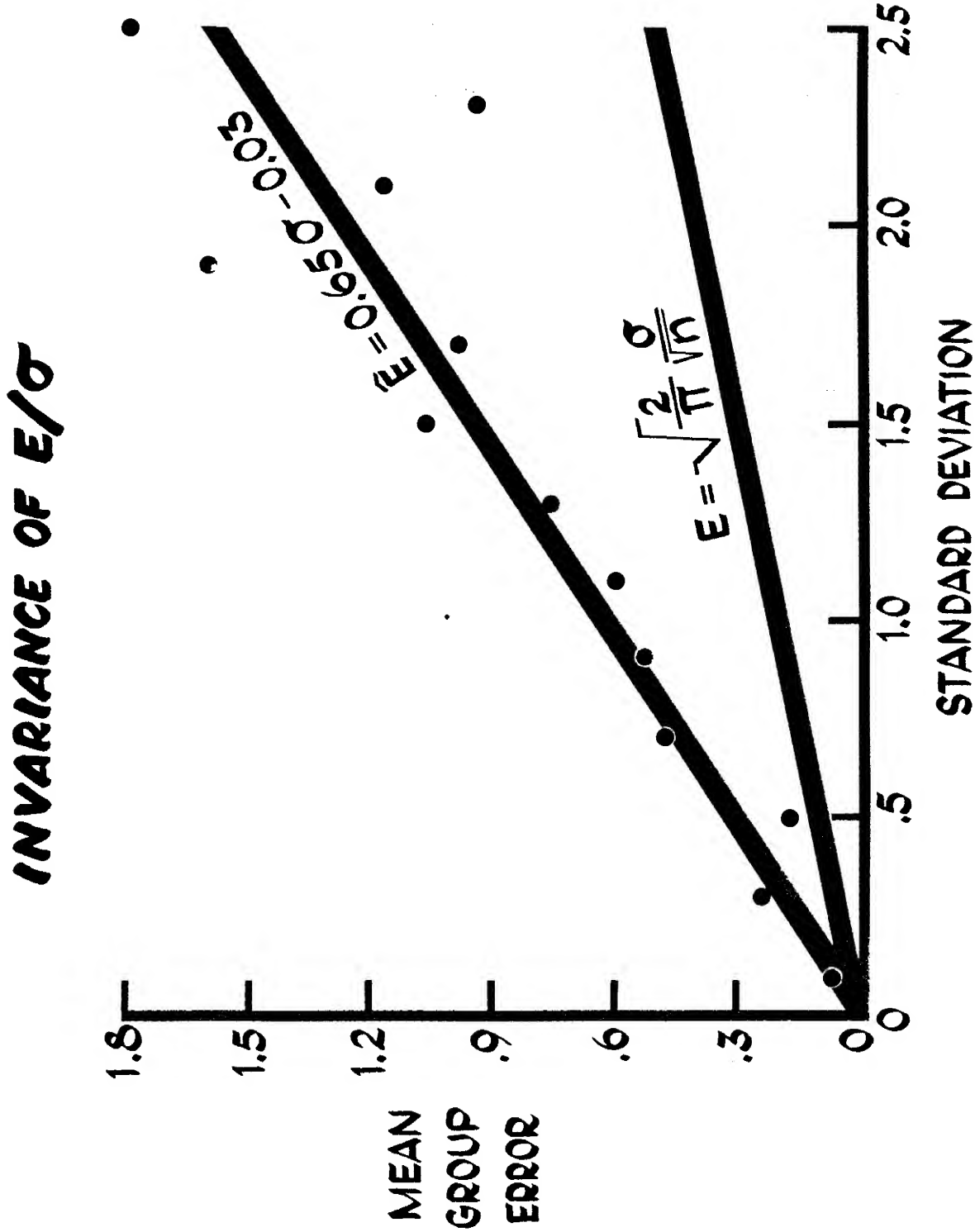
INVARIANCE OF E/σ
MEAN GROUP ERROR
STANDARD DEVIATION
1.8
1.5
1.2
.9
.6
.3
0
.5
1.0
1.5
2.0
2.5
E = 0.969σ - 0.023
E = √(2/π) σ/√n

Fig. 8

straight line (least squares fit). The small constant, .03, is probably within the noise level; and it appears reasonable to assume that the line passes through the origin. The lower line indicates the expected relation between average error and dispersion, assuming that estimation is a pure sampling process from a distribution centered on the true answer (for a sample of 14). The discrepancy between the observed and the expected error shows that the responses contain an appreciable amount of bias in addition to sampling error.

If we express the bias as the ratio E/σ, then the least-squares line in Fig. 8 indicates that the bias is a constant. This result appears to be highly significant for interpreting the results of the experiments.* In the first place, it substantiates the intuitive belief that higher dispersions are associated with decreased accuracy in a more dramatic fashion than the modest correlation mentioned above. More to the point, the fact that the bias is a constant suggests that estimation (in the absence of complete information) is a relatively well-defined process. The fact that the bias is greater than would be expected from a sampling process suggests that the discrepancy is a crude measure of the information deficiency in the estimates.**

---

*Finding an invariant of a process is always a refreshing, if somewhat rare, experience.

**This presumption is compatible with the results of Campbell. The short-range forecasting of economic indices is an information-rich task compared with our general information questions. If his round-one data are plotted on the $E-\sigma$ graph, they lie on a line well below the experimental line we obtained, and only slightly above the sampling error curve. However, he was dealing with too few cases to furnish a statistically significant relationship. An interesting possibility is that in asking for point estimates, we are doing something more like sampling means of individual distributions. In this case, we would expect a smaller standard deviation (and hence greater apparent bias) than if we had been sampling the distributions. This hypothesis receives some weight from the analysis of synthetic distributions described in Section 9.1, p. 50ff.

Figures 6 and 7 do not involve the accuracy of the responses. Figure 9 shows the distribution of individual scores where a subject's score is defined as the sum over twenty questions of the absolute values of the natural logarithms of his answers divided by the true answers. The distribution of second-round scores is also plotted in Fig. 9. The range of scores is again very large. For individuals at the far right of the curve, the average answer is off by a factor of about 8. At the low end of the curve, on the other hand, the subjects were off by an average factor of about 1.6.

It is natural to assume that such very large differences in scores on twenty questions indicate a wide range of ability to estimate. However, considering the large range of answers indicated by Fig. 6, a large range in scores would be expected to occur by chance. A crude test of the hypothesis that more than chance is operating is simply to take the actual answers on each of the questions and, so to speak, deal them out at random to respondents and compare the distribution of scores so obtained with the observed distribution. This has been done in Fig. 10, where the computed "randomized" scores are displayed in the smooth curve. It is clear from the figure that the randomized distribution is a relatively good fit to the actual, at the low end, but has a higher peak and a lower tail on the upper end. It would appear that more than chance is involved in the higher (lower accuracy) scores, and possibly also at the extreme low (high accuracy) end. We conclude that differences in ability to estimate the answer to general-information-type questions exist, but that these differences are heavily masked by chance.

Split-half reliabilities (correlation between accuracy on odd and even questions) for the first round range from about .4 to .6. This is not as good as would be desired for a measuring instrument. But the reliabilities do add to the presumption that differences in capability exist.

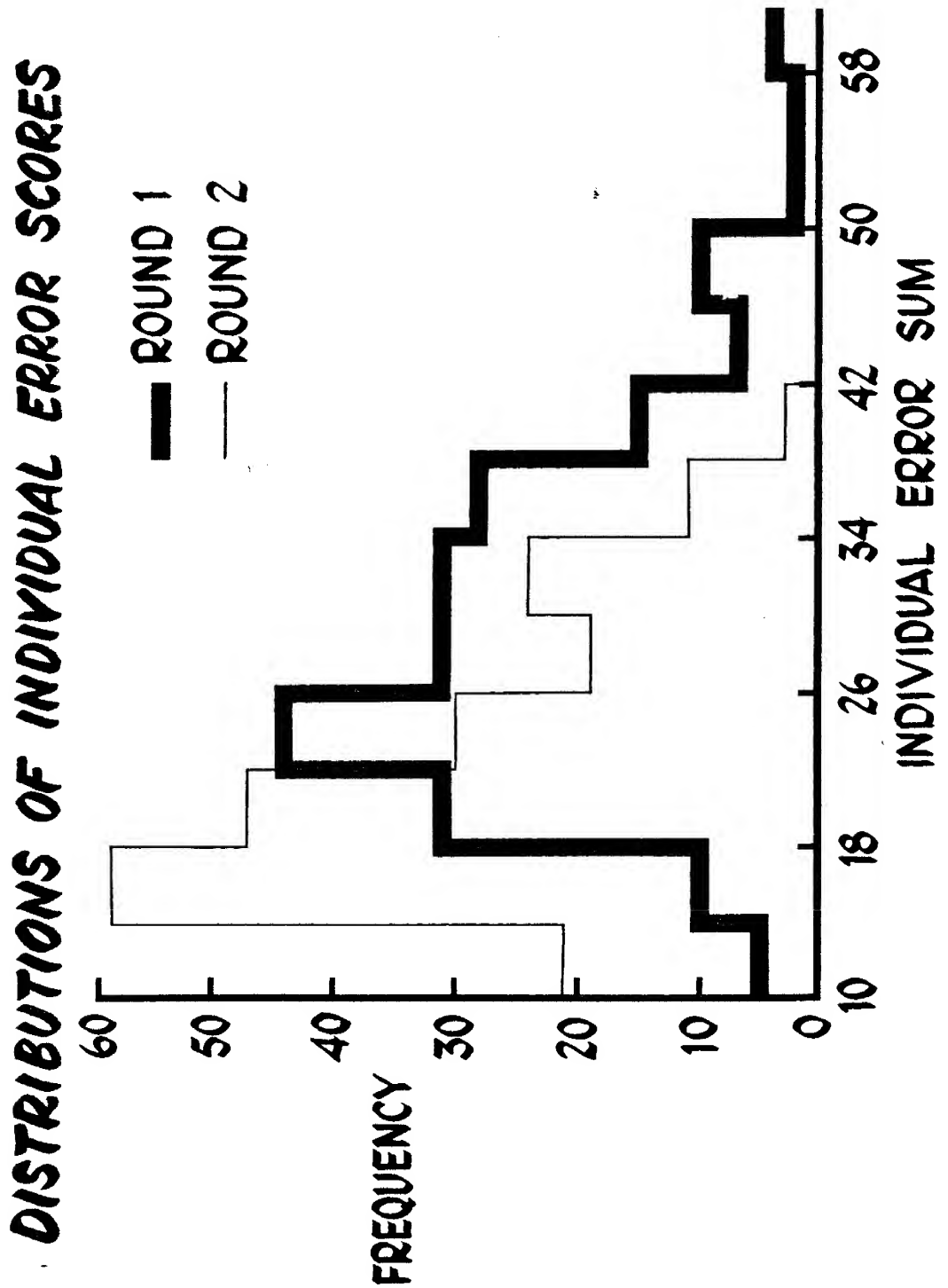
DISTRIBUTIONS OF INDIVIDUAL ERROR SCORES
ROUND 1
ROUND 2
FREQUENCY
0
10
20
30
40
50
60
10
18
26
34
42
50
58
INDIVIDUAL ERROR SUM

Fig. 9

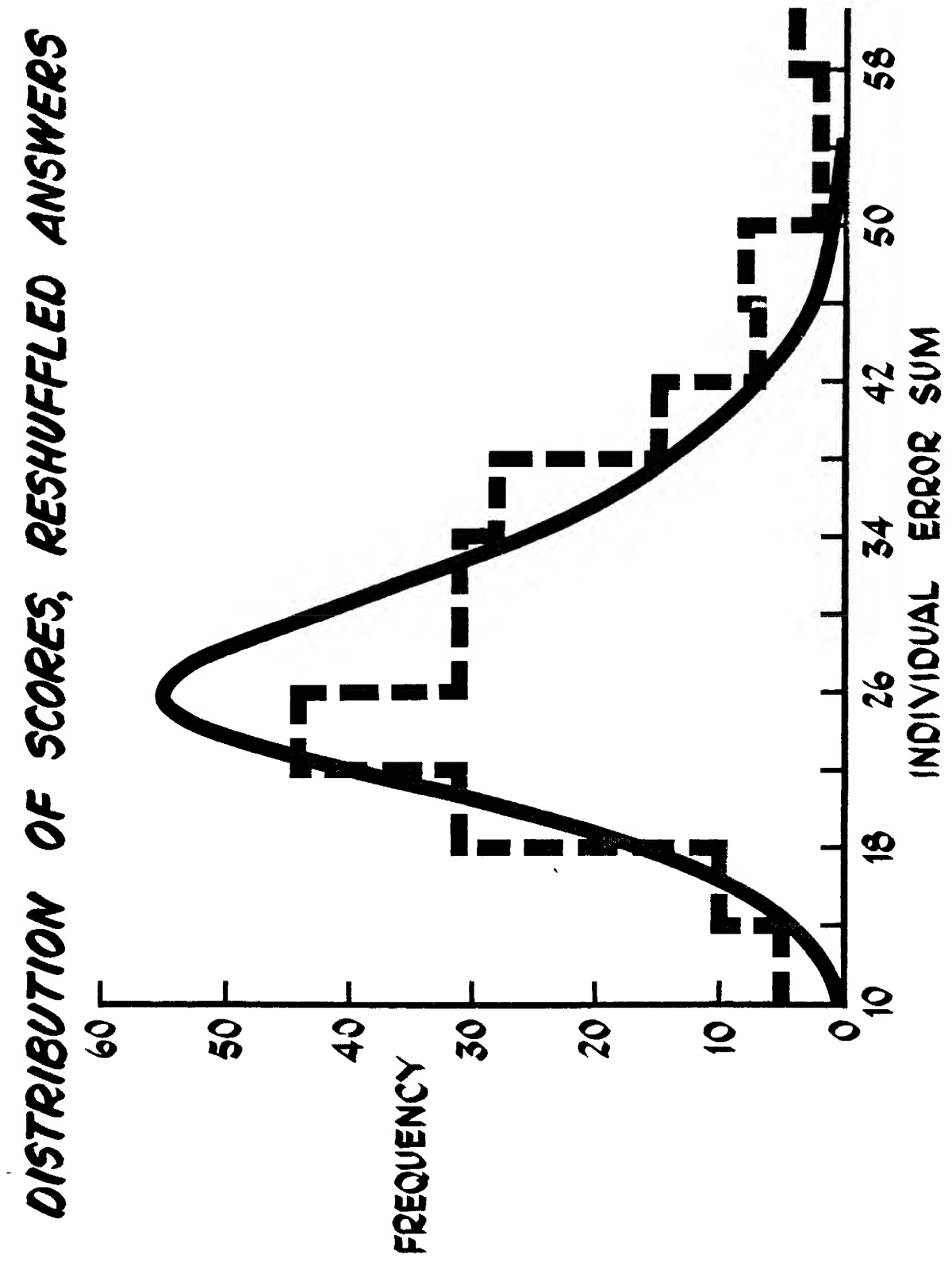
DISTRIBUTION OF SCORES, RESHUFFLED ANSWERS
FREQUENCY
0
10
20
30
40
50
60
10
18
26
34
42
50
58
INDIVIDUAL ERROR SUM

Fig. 10

A relatively extensive search for correlates of this presumed capability give a somewhat complex picture. This will be taken up in a later section.

The change from round one to round two shown in Fig. 9 indicates a large improvement in individual scores on iteration. Much of this change must be ascribed to convergence, i.e., to individuals whose first-round answers were highly divergent from the group median and who improved by moving toward the median.

7. IMPROVEMENT WITH ITERATION

The data presented in the previous section mainly concern the estimates of individual respondents. When we turn to group responses (defined as the median, or for some analyses, the geometric mean of the individual responses), the picture is pretty much the same, but with significant differences in degree. Figure 11 presents a cumulative distribution of group responses on the first and second round for 287 questions. In this case, the abscissa is the error rather than the deviation from the mean. A cumulative distribution is used in this instance, because the data are somewhat skimpy for a frequency distribution. Both curves are roughly log-normal. It can be seen by inspecting the two distributions that the second-round cumulative frequencies are uniformly above those for the first round. In short, on the second round, there is a higher proportion of second-round answers with lower errors; the second-round answers are to this extent better than the first-round answers.

The second observation with regard to the distributions is that the differences between them are small. The iteration step effected an improvement in accuracy, which was, however, less dramatic than the amount of convergence.

Table 2 presents the data on changes with regard to individual questions. Here the story is straightforward.

Table 2

IMPROVEMENT WITH ITERATION AND FEEDBACK

	Number of questions
More Accurate	89
Same	80
Less Accurate	51

For about 64 percent of the changed estimates, the median improved in accuracy; for 36 percent, the median became less accurate. Perhaps more impressively, in none of the eleven

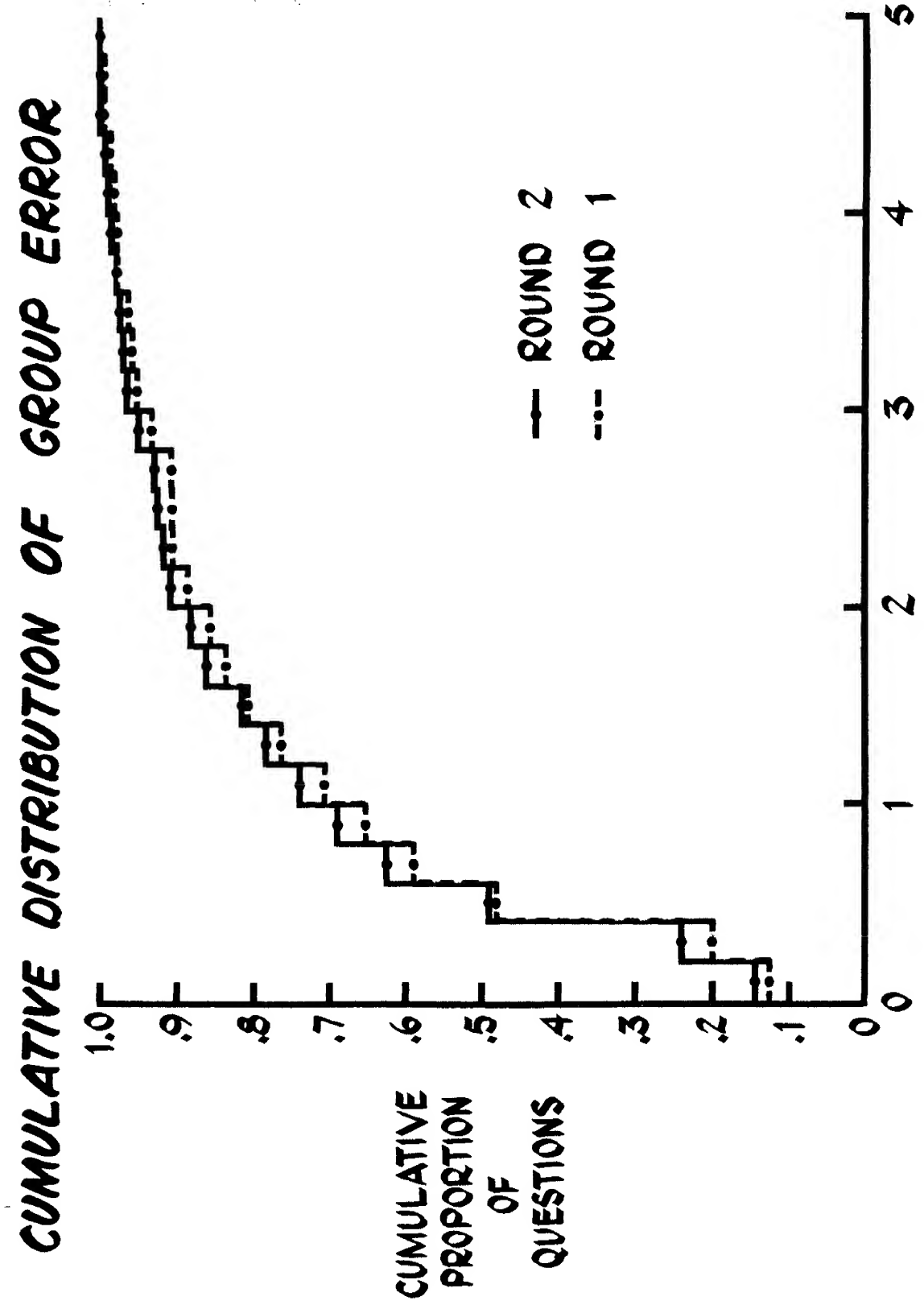
CUMULATIVE DISTRIBUTION OF GROUP ERROR
CUMULATIVE PROPORTION OF QUESTIONS
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
1
2
3
4
5
ABSOLUTE GROUP ERROR
— ROUND 2
-- ROUND 1

Fig. 11

individual groups represented in Table 2 did the number of decreases in accuracy exceed the number of increases.*

This is, so to speak, the basic outcome of the series of experiments and furnishes the basis for presuming the Delphi procedures to be useful. However, as was evident from Fig. 11, the improvement between round one and round two is not particularly impressive. The question arises whether it is possible to improve on this result. To do so, it seems likely that a deeper understanding of the mechanism of improvement is necessary.

---

*For the grouped data in Table 2, considered as a process with a single degree of freedom (under the hypothesis that a decrease in accuracy is as likely as an increase), $\chi^2 = 10.3$, $p < .01$. When the data are analysed in terms of the individual groups, using a trinomial distribution on better, same, and worse, and assuming that the likelihood of same is the proportion experimentally found, 4/11, the probability of outcomes as good or better than the experimental is .003.

8. MECHANISM OF IMPROVEMENT

In order to have improvement, there must be changes of estimates between rounds. The most obvious influence producing change is the feedback of the round-one median. As Fig. 12 indicates, most of the tendency to change can be ascribed to one parameter, namely, the distance of the first-round answer from the first-round median. (The abscissa is in terms of distance from the median measured in units of the upper quartile minus the median for answers above the median, and in units of the median minus the lower quartile for answers below the median.) The likelihood of a change of estimate is very nearly a linear function of the distance from the median to about two quartiles, at which point it becomes erratic in a rather charmingly symmetrical fashion. Some of the erratic behavior beyond 2 can be ascribed to small samples; but some probably should be ascribed to the effect of feeding back the quartiles. The number of individuals who changed directly to the nearest quartile is much larger than would be expected from simple movement toward the median. In addition, we would expect that individuals who are close to the quartiles would be less likely to change than those who are far away from any reference point. There were not enough data to examine this hypothesis in detail.

Movement toward the median is not enough in itself to account for change in the median between round one and round two. In order for any change in the median to occur, it is necessary that some respondents make changes that cross the median. This can be clearly seen by assuming that all subjects simply move to the median. Although the degree of convergence would be as high as possible in this case, the median would not change. A fortiori, in order to effect a systematic improvement in the median, some process beyond simple convergence must be in operation.

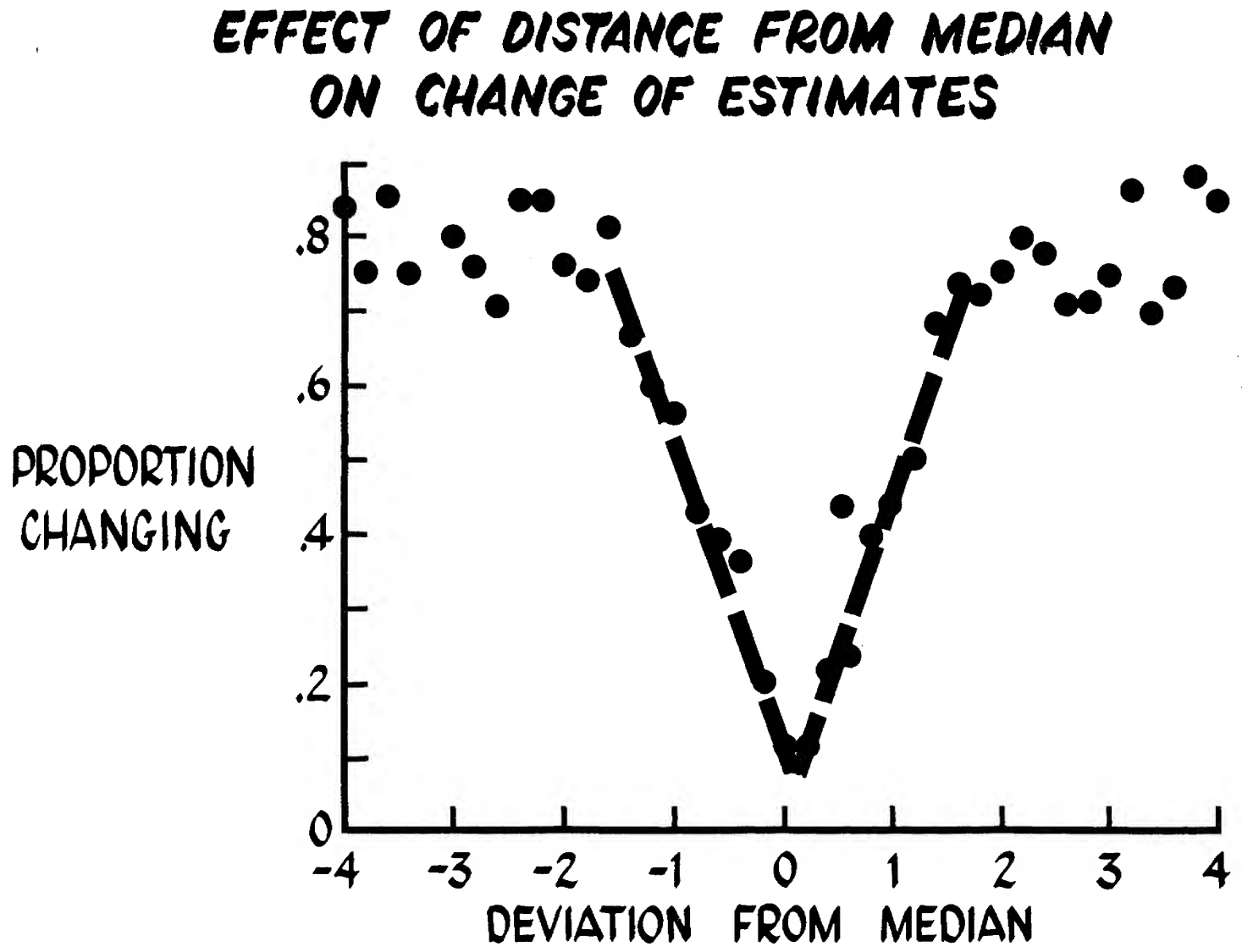
EFFECT OF DISTANCE FROM MEDIAN
ON CHANGE OF ESTIMATES
PROPORTION
CHANGING
.8
.6
.4
.2
0
-4
-3
-2
-1
0
1
2
3
4
DEVIATION FROM MEDIAN

Fig. 12

It is convenient to divide the group (for a given question) into the holdouts—those who do not change their estimates at all from round one to round two—and those who do change, the swingers. Figure 12 showed that the holdouts tend to cluster about the median. It is also the case that the holdouts tend to be more accurate in their first-round estimates than the swingers. Table 3 presents the comparison between the accuracy of the holdouts, swingers, and total group for round one and for round two. It is also evident from the table that the holdouts are more accurate than the total group on round one.

Table 3

MOST ACCURATE SUBGROUP
(Geometric Means)*

	Round 1		Round 2	
Holdouts	141		113	
Swingers	73	94[a]	101	125[a]
Total Group	21		24	
Ties	5		2	

[a]The total group is more accurate than the holdouts in the bracketed cases.

Figure 13 illustrates in schematic fashion the situation in round one and the effects of convergence. The mean of the group M_G will always lie between the mean of the swingers M_S and the mean of the holdouts M_H. Since the mean of the

*Table 3 is based on geometric means, rather than medians, to allow the differences to stand out more clearly. Although the geometric mean was slightly less accurate than the median, the geometric mean exhibited more changes between round one and round two.

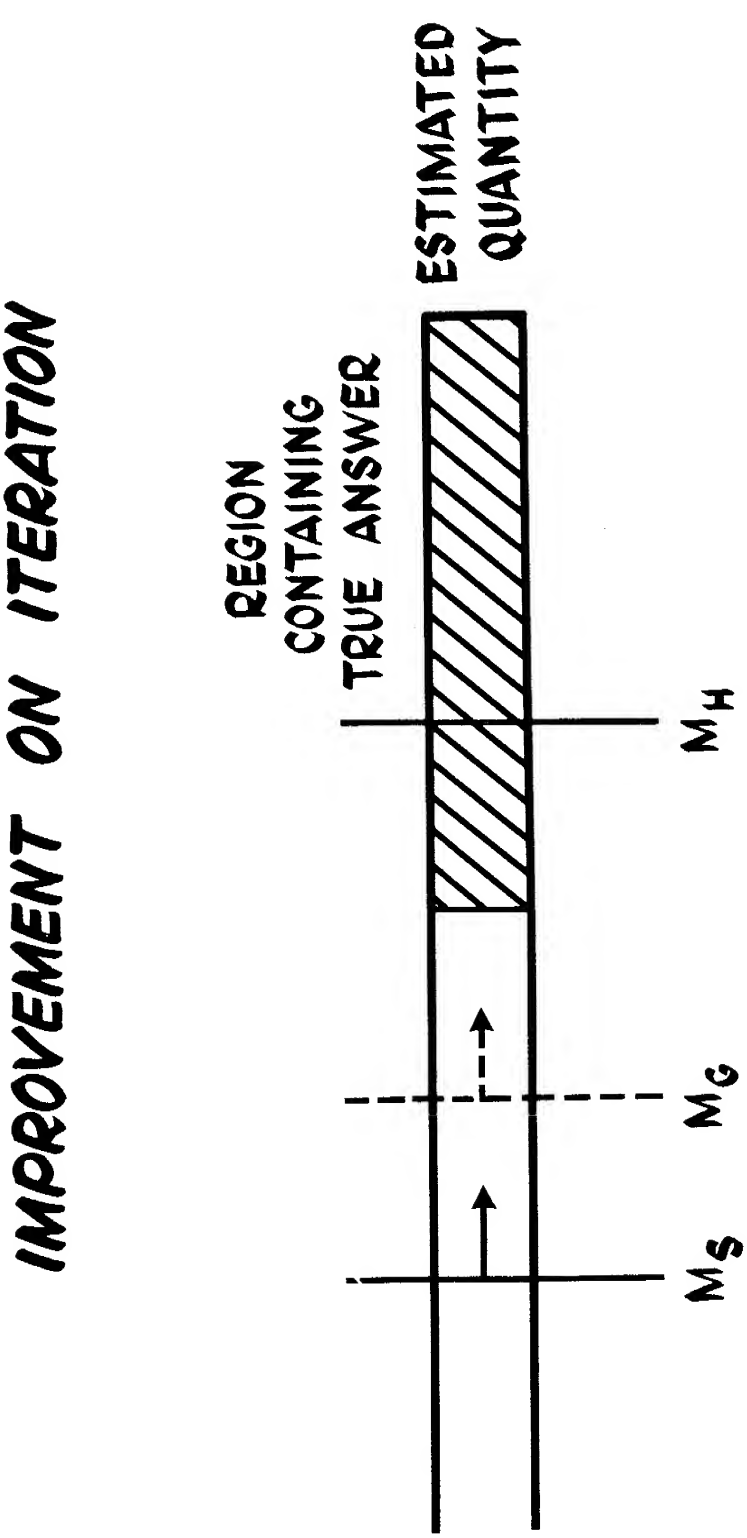
IMPROVEMENT ON ITERATION
ESTIMATED QUANTITY
REGION CONTAINING TRUE ANSWER
M_H
M_G
M_S

Fig. 13

holdouts is closer to the true answer than either the mean of the swingers or the mean of the group, the true answer must lie somewhere in the shaded region. It is immediately clear from the figure that if the mean of the swingers moves in the direction of the mean of the group, the mean of the group will also move to the right and, in general, will improve. It will become less accurate only in case it moves across the mean of the holdouts, which requires that the mean of the swingers also move across the mean of the holdouts. If we define convergence as movement toward the mean on the part of these who move, then the kind of degradation mentioned cannot occur with convergence alone.

To sum up the preceding: A first approximation to a model of improvement on iteration is afforded by two assumptions: (1) the holdouts are more accurate than the swingers and than the total group on round one; (2) on iteration, the mean of the swingers moves toward the mean of the total group. These two assumptions are sufficient to assert that the mean of the total group will improve.

The first approximate model is insufficient to explain why the total group is more accurate than the holdouts on round two. The disparity is not sufficient to make much of in itself; however, the shift from 94 to 125 cases in which the total group is more accurate than the holdouts is inexplicable on the approximate model. In case the holdouts are more accurate than the total group, and convergence is the only process occurring, then the holdouts will remain more accurate than the total group. The approximate model, then, is an explanation for part of the improvement in the mean of the total group on iteration; but a significant amount of improvement remains to be explained.

In Fig. 14 the average <u>amount</u> of change is plotted as

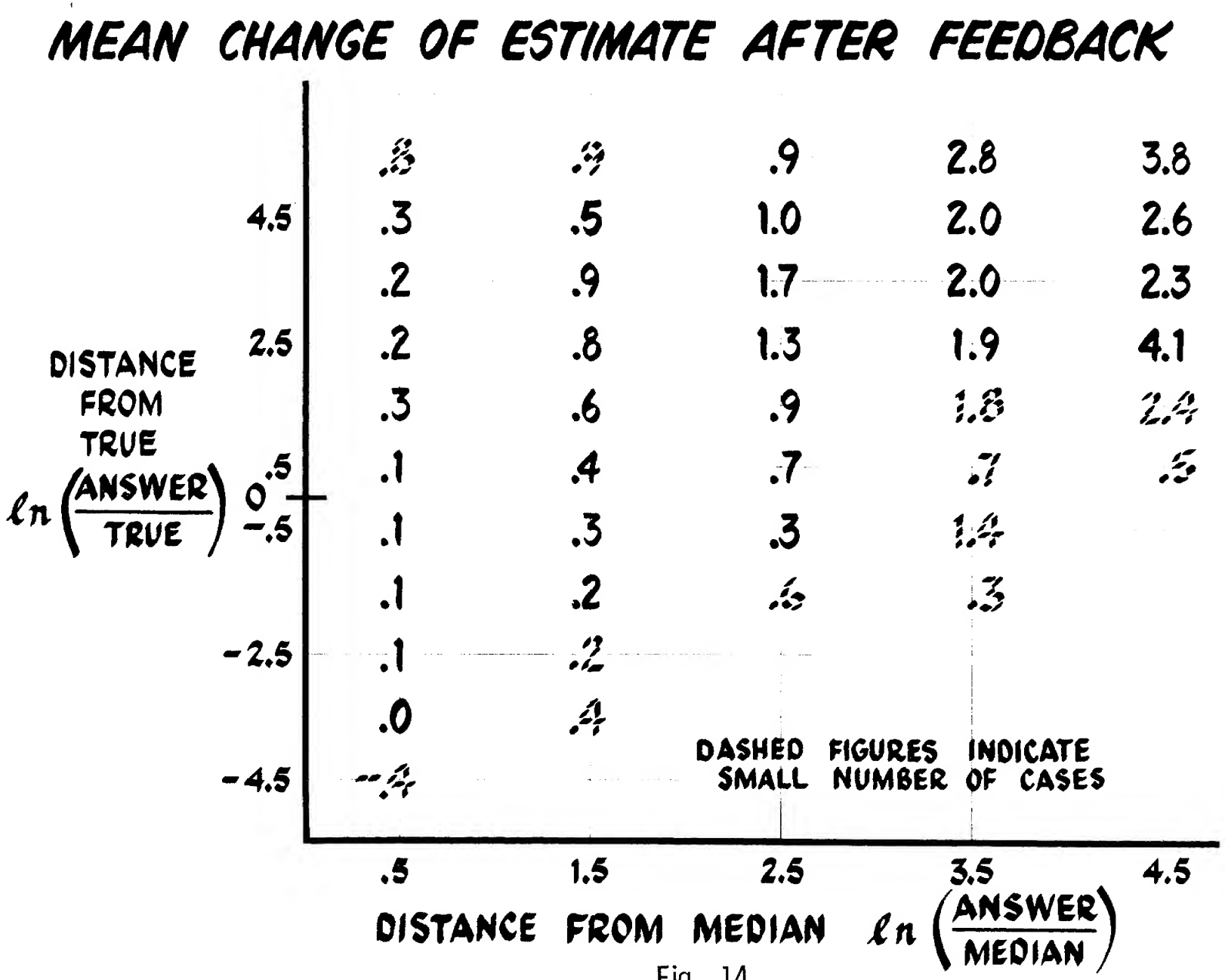
MEAN CHANGE OF ESTIMATE AFTER FEEDBACK
DISTANCE FROM TRUE ln (ANSWER / TRUE)
4.5
2.5
0
.5
-.5
-2.5
-4.5
.8 .9 .9 2.8 3.8
.3 .5 1.0 2.0 2.6
.2 .9 1.7 2.0 2.3
.2 .8 1.3 1.9 4.1
.3 .6 .9 1.8 2.4
.1 .4 .7 .7 .5
.1 .3 .3 1.4
.1 .2 .6 .3
.1 .2
.0 .4
-.4
DASHED FIGURES INDICATE SMALL NUMBER OF CASES
.5 1.5 2.5 3.5 4.5
DISTANCE FROM MEDIAN ln (ANSWER / MEDIAN)

Fig. 14

a function of two variables, the distance of the first-round answer from the median and the distance of the first-round answer from the true answer.* Distance is measured by the logarithm of the answer divided by the median in the first case and by the true answer in the second case; amount of change is measured by the average change in log scores in the appropriate box. We have already seen that distance from the median has a very strong influence (Fig. 12) with respect to the likelihood of change; Fig. 14 shows that the distance from the median has an equally strong effect on the amount of change.

It is clear from Fig. 14 that the effect of the median is much stronger than the effect of the true answer, almost to the extent that the median effect completely dominates the effect of the true answer within the region bounded by 3.5 on each axis. On the other hand, the effect of the distance from the true is evident. This is brought out more clearly in Fig. 15, where the amount of change is plotted against distance from the true for two constant deviations from the median. The curves (approximated by hand) indicate a definite increase in motion with distance from the true answer.

In a crude, but illuminating, analogy with a physical model, we can speak of two forces operating on a subject to bring about a change of opinion. One force, which is a function of the distance of the subject's answer from the median, tends to change the opinion on the second

---

*In order to obtain sufficient cases for the entries, the two left quadrants have been reflected into the right quadrants, with quadrant II reflected into quadrant IV and quadrant III reflected into quadrant I. The assumption of symmetry was based on Fig. 12 and inspection of the full plot. The dashed entries indicate cases for which the number of instances were insufficient to give a reliable estimate.

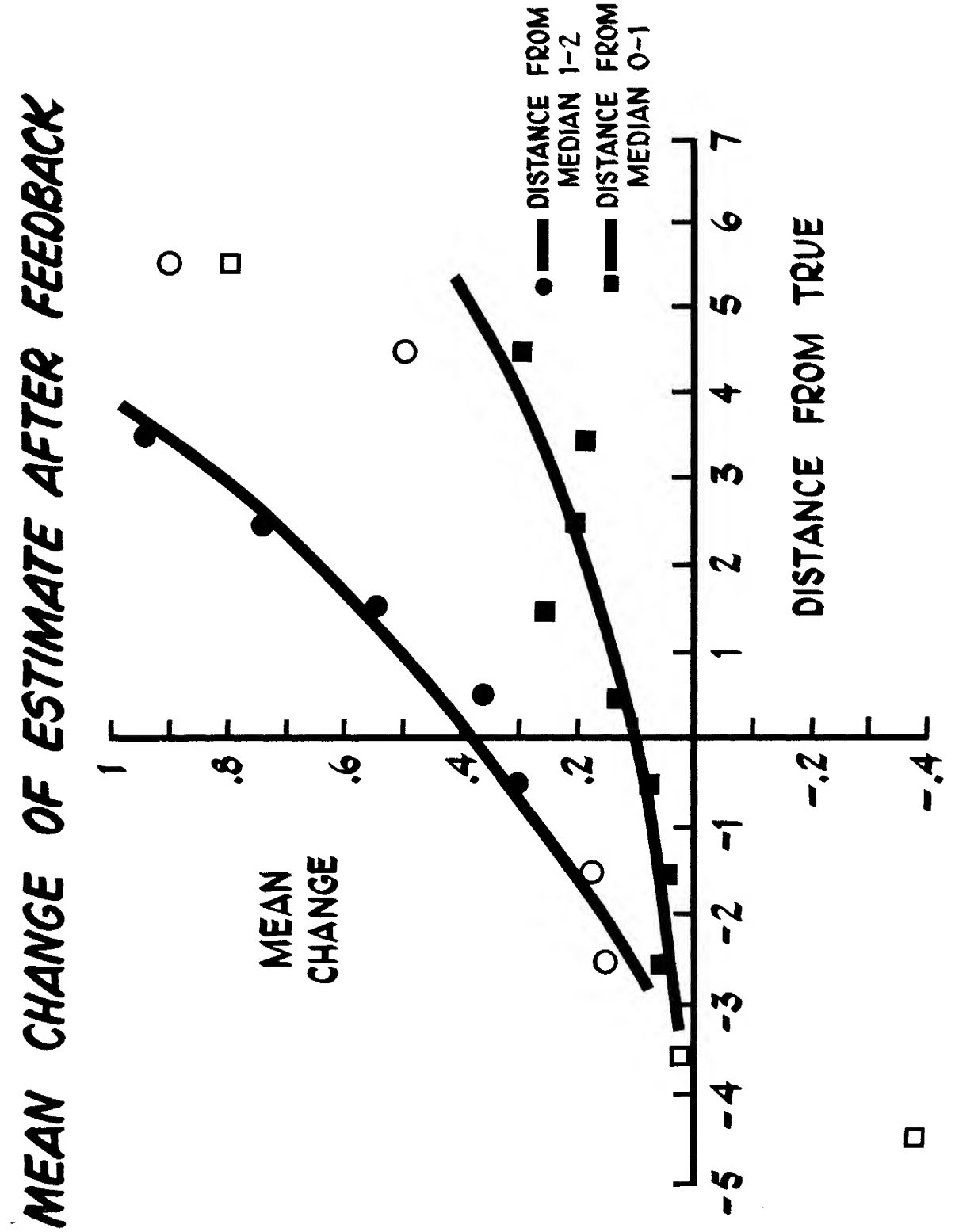
MEAN CHANGE OF ESTIMATE AFTER FEEDBACK
MEAN CHANGE
1
.8
.6
.4
.2
-.2
-.4
-5 -4 -3 -2 -1 1 2 3 4 5 6 7
DISTANCE FROM TRUE
DISTANCE FROM MEDIAN 1-2
DISTANCE FROM MEDIAN 0-1

Fig. 15

round in the direction of the median. The other force, which is a function of the distance of the subject's answer from the true answer, tends to move the opinion toward the true. The "pull of the median" is much stronger than the pull of the true, but both operate.

In another way of speaking, it would appear that there is a certain amount of residual information remaining in the group after the first-round estimates have been expressed. In a fashion not yet explicable in terms of our data, the iteration and feedback step causes (or allows?) this additional information to be brought into play, with consequent improvement in the group estimate.

On the analogy with physical forces, the pull of the true answer is desirable. It has been thought in the past that the pull of the median is also desirable on the grounds that it leads to convergence and greater agreement among the respondents. In part this is due to the presumption discussed in Section 6 that greater agreement implies greater accuracy. In part it is probably also influenced by several subsidiary issues—namely, it is easier to use an estimate with a narrow spread (how to take into account the uncertainty expressed by a wide spread?), and if the concurrence of the group in some decision is required, greater acceptance of the decision would occur if opinions were fairly close.

The two practical problems—reduction of "uncertainty" and concurrence on decisions—need further study. But the presumption that greater agreement implies greater accuracy needs modification when the agreement results from convergence. Figure 16 shows the average error as a function of standard deviation for both round one and round two. The round two data are not as neatly linear as those for round one; but the important lesson from Fig. 16 is that the bias is greater on round two than on round one. There is good

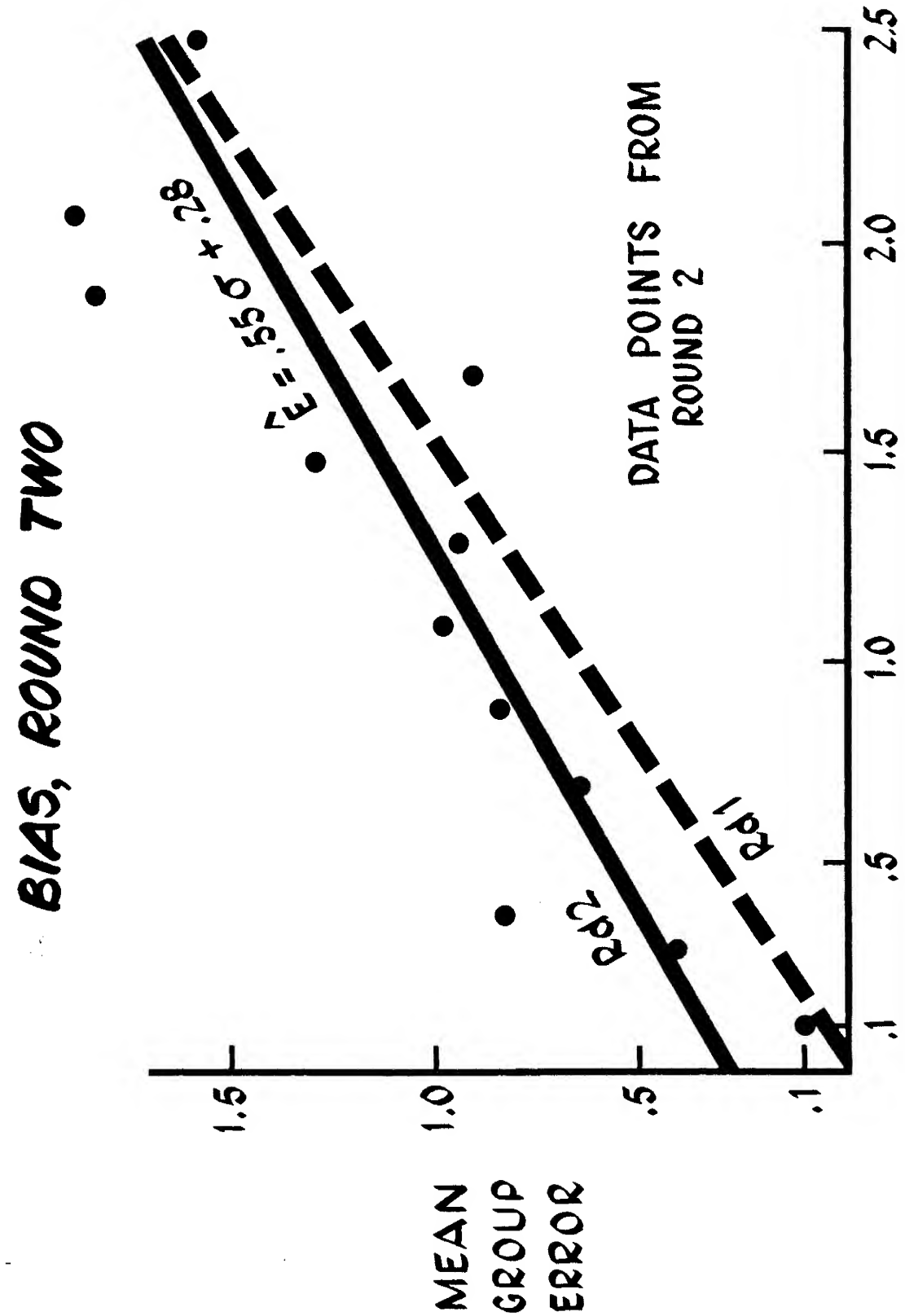
BIAS, ROUND TWO
MEAN GROUP ERROR
STANDARD DEVIATION
DATA POINTS FROM ROUND 2
1.5
1.0
.5
.1
2.0
2.5

Fig. 16

reason to assert that too much convergence has occurred; the increase in accuracy is not commensurate with the reduction in spread.

The question is; Can the pull of the true be amplified and the pull of the median be damped? There are some grounds for optimism. One possible approach is to try feeding back something weaker than the three quartiles. As an example, feedback might be individualized for each respondent, stating his percentile in the round-one distribution. This would remove the median as a sharp "target" for his changed estimate and at the same time maintain the motivation to change. This possibility will be tested in further experiments.

Figure 17 displays the amount of change of the group response as a function of group error. As might be expected, the amount of change increases monotonically with the amount of error. Also, as might be expected, the average change becomes negative (i.e., represents a net motion away from the true answer) as the initial group error becomes small; in effect, when the group is very accurate, any change is likely to be for the worse.

Figure 17 should be taken into account in assessing the significance of the earlier statement (p. 34) that the amount of improvement of the group response on iteration is small. For a large proportion of our questions, the initial error was small, and hence changes were small.

A somewhat more interesting question is whether this result, in combination with other results concerning the accuracy of group responses, can be exploited to improve the Delphi procedures. In general, it would appear that if the more accurate responses on round one could be identified, then with present procedures it might be better to omit the iteration step for those questions. Potential techniques for assessing the accuracy of responses will be discussed in Section 9.8, p. 68ff.

CHANGE OF GROUP ESTIMATE AS A FUNCTION OF GROUP ERROR

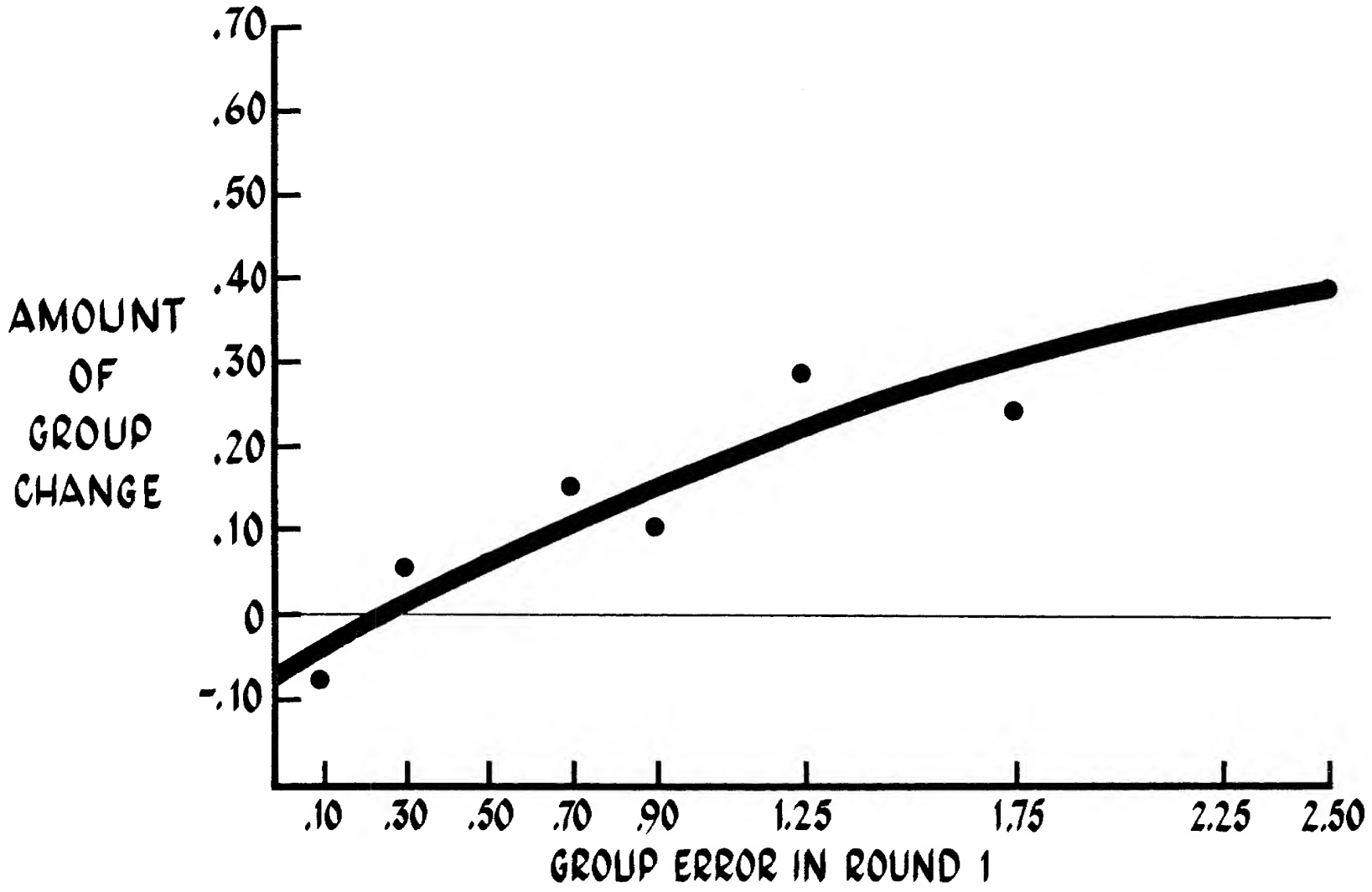

Fig. 17

9. SUPPLEMENTARY ANALYSES

In addition to the basic issues discussed above, the experiments generated data concerning a number of other pertinent features of group estimation. The items presented below represent a selection of the more interesting results from these substudies. Other material will be reported in later publications.

9.1 Distributional Estimates. A plausible hypothesis with respect to opinion is that estimators have "in the back of their minds" a rough probability distribution over the quantity in question; and when requested to produce a single (point) estimate, they select some measure of central tendency for this distribution. If this is the case, then theoretically, a more accurate estimate could be obtained by summing the individual distributions and selecting the mean or median of the composite distribution as the group response.

Two experiments were devoted to examining the effect of requesting distributional responses rather than point estimates. Subjects were asked to furnish the three quartiles for each question, that is, the number for which there is a 25 percent chance that the true answer is less, the number for which there is a 50 percent chance the true answer is less, and the number for which there is a 75 percent chance that the true answer is less. The three were called the low, mid, and high estimates, respectively. Somewhat to our surprise, the subjects had no difficulty making these presumably more complex estimates.

In the first experiment, there was a control group that made point estimates for the same questions. In the second experiment, there was no control group, and an evaluation can be made only by comparison with other groups in the series. In the first experiment, the experimental group

was more accurate on both rounds, as shown in Table 4, the difference being heightened on round two. The second half of Table 4 shows the improvement for the experimental and control groups between rounds. The experimental group demonstrated a greater improvement. Neither of these two results is statistically significant by themselves.

Table 4

COMPARISON OF DISTRIBUTIONAL VS POINT ESTIMATES

	Round 1	Round 2
More Accurate	10	12
Same	2	0
Less Accurate	8	8

IMPROVEMENT BETWEEN ROUND 1 AND ROUND 2

	Experimental	Control
More Accurate	14	7
Same	1	12
Less Accurate	5	1

In the second experiment (30 subjects), the median improved in 10 cases, remained the same in 9, and became less accurate in only 1 case. Hence, the amount of improvement between rounds was much greater than for any other experiment in the series.

There is one consideration that clouds the results for the first experiment somewhat: Rather than feeding back medians and quartiles, the means of the three individual quartiles were fed back. It happens that for this particular group, the members tended to underestimate in most of their answers. As a result, the mean tended to be more accurate than the median. There is no way without further experiments to determine how greatly the improvement was dependent on this fact. In the second experiment, the median and

quartiles of the mid estimates were fed back.

Group distributions were constructed in the following way: Individual distributions for each question were approximated by constructing a triangle on each side of the mid estimate so that one-half of the area of the triangle was included by the section between the mid and quartile estimates (see Fig. 18). These individual distributions were then summed to produce a group distribution. An example of such a synthetic group distribution is shown in Fig. 19.

Examination of the two sets of synthetic distributions indicated that the mode was a somewhat more accurate estimator than the mean. Accordingly, the mode was used for the group response to compare with the medians of the point estimates. For the first group, the mode of the calculated distribution was more accurate than the median of the mid estimates for 12 cases and less accurate in 8 cases. However, for the second group, the mode of the calculated distribution was more accurate in only 7 cases and less accurate in 10, with 3 ties. Considering the crudeness of the approximations used, these results are quite encouraging and suggest that more careful ways of deriving a group distribution may result in more accurate estimates.

A highly suggestive finding is that the dispersion of the synthetic distributions is much greater than the dispersion of the distribution of point estimates. The average ratio of the standard deviations of the synthetic distributions to the standard deviations of the point estimates is 3.0. If the abscissa scale for the experimental data of Fig. 8 is expanded by a factor of three, the experimental curve lies almost on top of the theoretical sampling error curve. The apparent large bias of the experimental data may be the result of underestimating the true dispersion of the "underlying" distributions in the minds of the respondents.

SYNTHETIC DISTRIBUTION WITH LOW, MID, HIGH ESTIMATES

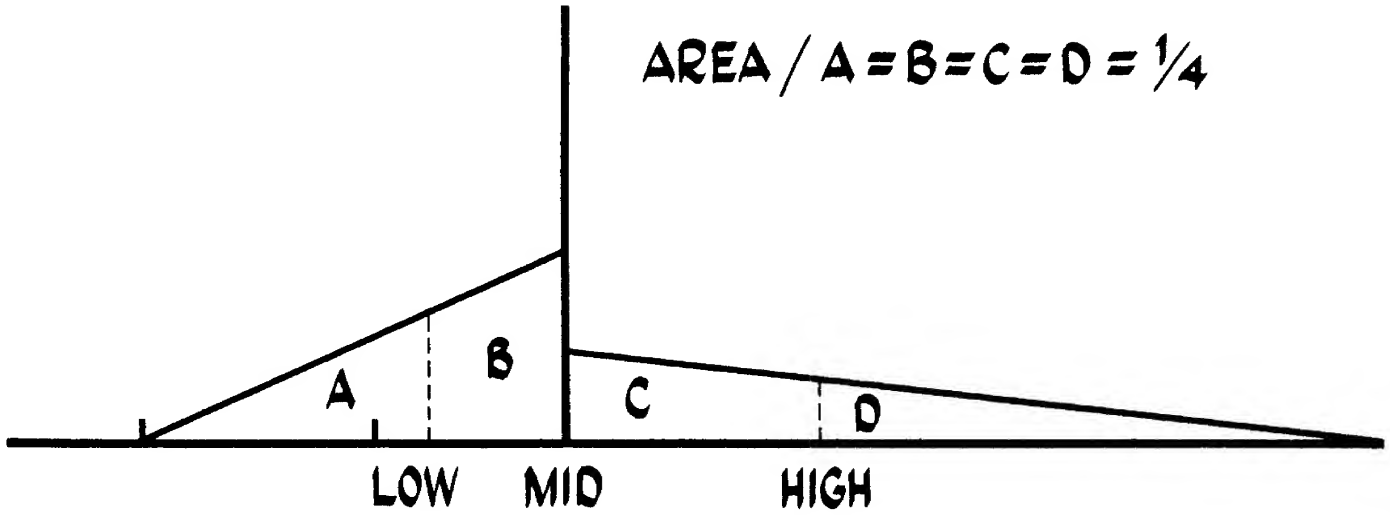

Fig. 18

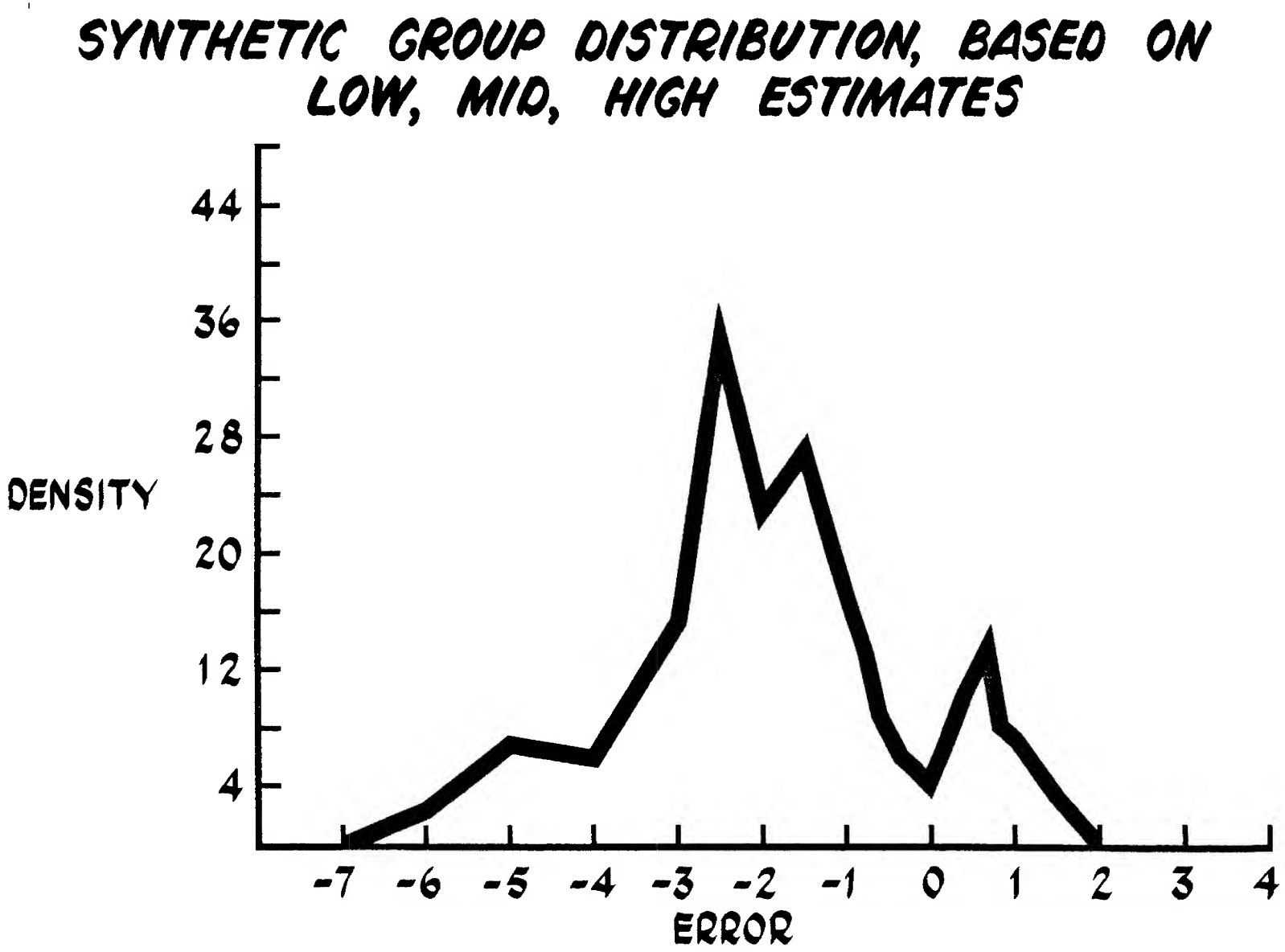
SYNTHETIC GROUP DISTRIBUTION, BASED ON LOW, MID, HIGH ESTIMATES
DENSITY
44
36
28
20
12
4
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
ERROR

Fig. 19

9.2. Learning. A potential criticism of the procedures being investigated with regard to applications is that the subjects were fairly "naive" in the task assigned, whereas in applications it would be expected that the respondents would be experts with long experience in making the kind of estimate involved in the exercise. The rather small amount of experimental data and much larger experience with non-experimental applications suggests that this is probably not the case, but the evidence is certainly not sufficient to give an unequivocal reply.

One consideration here is whether the estimation task is a skill that can be learned. We devoted one experiment to testing the hypothesis that it was. In this experiment, the questions were presented one at a time. The estimation, feedback, and reestimation were completed before going on to the next question. In addition, after completion of the iteration, the group was told its second-round median and the true answer. Thus, members of the group could compare both their own performance and the performance of the group, question by question. The results are shown in Figs. 20 and 21. Figure 20 shows the individual performance as a function of question order averaged over blocks of five questions. There is a clear downward trend in round one, indicating some learning. This effect does not show up in round two, where the improvement between the two rounds is greatest for the first (least accurate) block.

Figure 21 is a similar curve for group scores, and no discernible trend is indicated. Furthermore, no group improvement between round one and round two is discernible for the later blocks of questions. We thus have no evidence for group learning with a sequence of twenty questions.

These results indicate that, although a discernible learning effect exists for individual responses, this effect is dominated by the effects of feedback and aggregation into a group response.

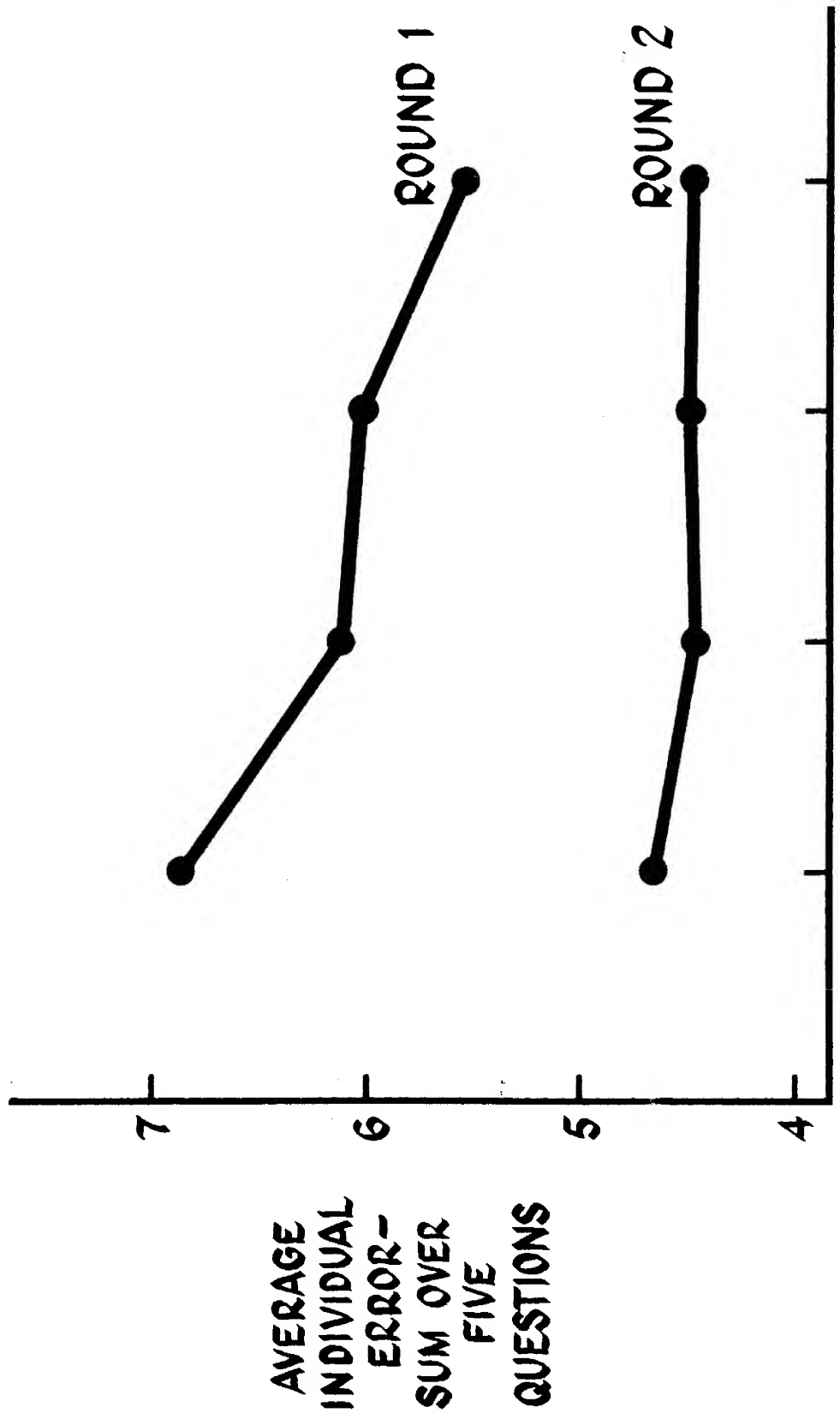
INDIVIDUAL LEARNING
AVERAGE INDIVIDUAL ERROR- SUM OVER FIVE QUESTIONS
7
6
5
4
ROUND 1
ROUND 2
1-5
6-10
11-15
16-20
QUESTION NUMBER

Fig. 20

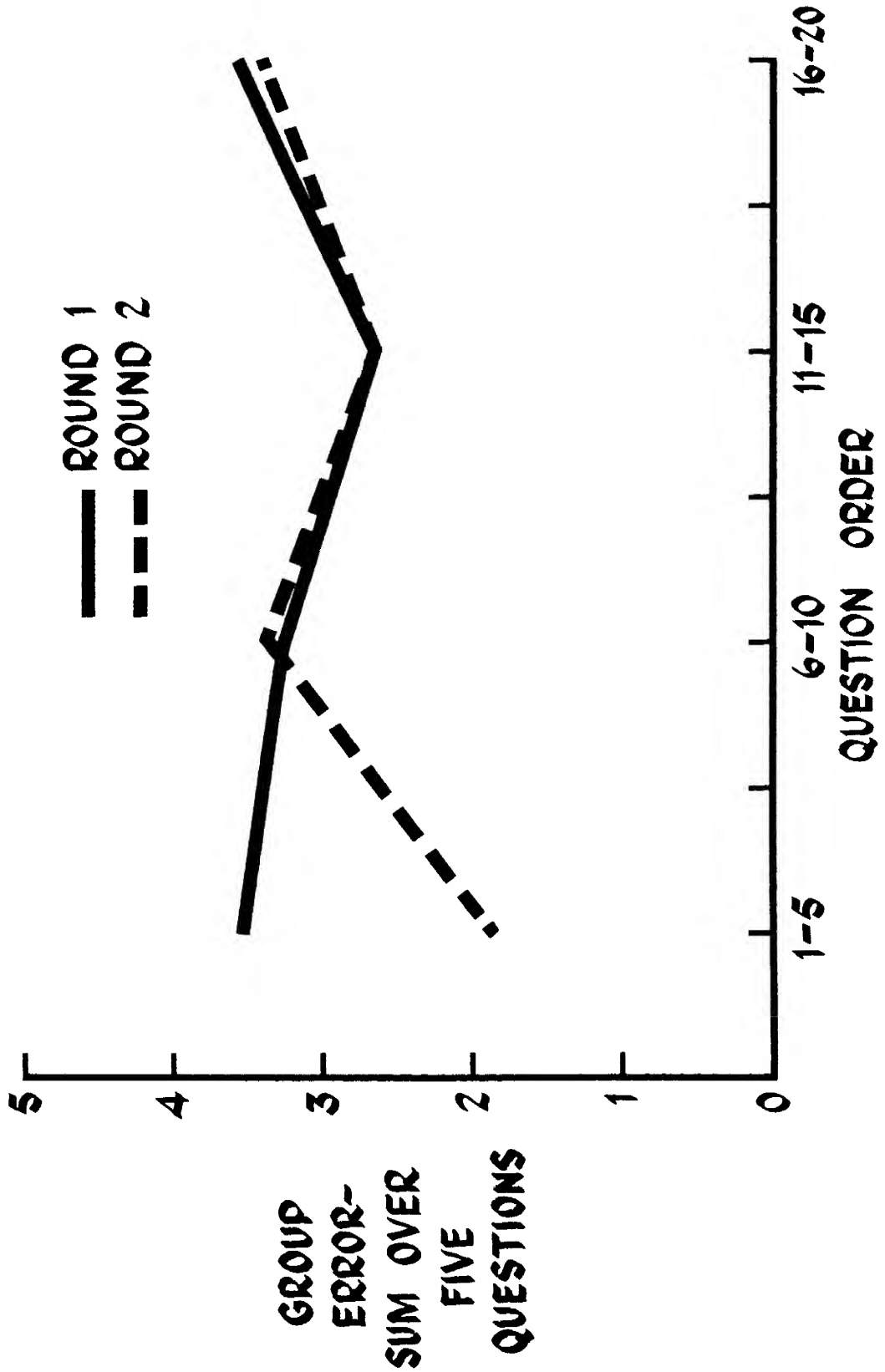
GROUP "LEARNING"
ROUND 1
ROUND 2
GROUP ERROR-SUM OVER FIVE QUESTIONS
0
1
2
3
4
5
1-5
6-10
11-15
16-20
QUESTION ORDER

Fig. 21

9.3. Other Forms of Feedback. It has been customary in applications [10] to include other types of feedback in addition to the statistics of the previous-round answers. A typical procedure is to ask the subjects who are at the two extremes (i.e.) in the top and bottom quartiles) on the first round to write down their reasons for their answers. These are edited by the exercise managers and fed back along with the second-round statistics on the third round. A fourth round may include the formulation of counterarguments.

In our first experiment comparing Delphi with face-to-face interaction, this procedure was followed in the Delphi sessions. There was no control group with respect to feedback of reasons. However, a highly suggestive outcome of this experiment was that the answers were most accurate on round two and became less accurate on subsequent rounds. Whether this deterioration can be ascribed to the feedback of reasons cannot be determined from the experiment; but we can conclude that there is no evidence that the reasons helped.

A second experiment was devoted directly to examining the effects of feeding back reasons, using a control group. The experimental group was instructed on the second round to formulate reasons for their opinion if their response on the second round was outside the interquartile range of the first round. These reasons were summarized and fed back on the third round in addition to the medians and quartiles of the second round. A control group underwent a similar set of three rounds, except that reasons were not asked for on round two but on round three. These were not fed back. The point of asking for reasons on round three for the control group was to determine whether the task of formulating reasons would have a discernable effect on the individual and group responses.

The outcome of this experiment is given in Table 5 .

Table 5

CHANGES IN GROUP RESPONSE WITH AND WITHOUT FEEDBACK OF REASONS

	Without Feedback of Reasons			With Feedback of Reasons		
	Better	Same	Worse	Better	Same	Worse
Between rounds one and two	4	12	4	6	8	6
Between rounds two and three	4	15	1	4	8	8
Between rounds one and three	6	10	4	8	1	11

Although none of the changes are significant in this experiment, we can say unequivocally that the addition of formulating and feeding back reasons did not increase the accuracy of initial estimates or produce greater improvement on iteration.

Two experiments were concerned with additional feedback of another sort. In this exercise the experimental group was asked to answer two related questions in addition to the primary question. Two hypotheses were being tested: (1) The task of responding to related questions would stimulate the subjects to consider a richer set of relevant factors, and thus improve accuracy. (2) The responses of the group on the related questions, when fed back as medians and quartiles, would act as additional information available to the subjects and hence increase accuracy. The outcome of the experiment was indecisive. Table 6 indicates the comparison of improvement for the two treatments.

Table 6

IMPROVEMENT ON ITERATION WITH AND WITHOUT RELATED QUESTIONS

	With Related Questions	Without Related Questions
Better	9	6
Same	6	11
Worse	5	3

ACCURACY COMPARISON BY QUESTION

Experimental Group was	Round 1	Round 2
Better	10	12
Same	8	6
Worse	2	2

In the second experiment, there was no control group. Related questions were asked for 10 of the 20 questions. In 4 of the 10 cases with related questions, answers improved on iteration, in 6 cases they became worse. For the 10 questions without related questions, there were 5 improvements, 3 worse, and 2 ties.

The two experiments give contrary indications with regard to the effectiveness of related questions. In the first experiment, related questions appear to improve the group performance both with respect to initial accuracy and with respect to improvement after feedback. In the second experiment (where the group was its own control), answering the related questions appears, if anything, to degrade the group performance. Probably somewhat more weight should be given to the first experiment, in which case the evidence is slightly in favor of including related questions. Additional experiments appear necessary before a firm conclusion can be reached.

9.4. Sexual Differences. The experiments verified two widely held clichés concerning the differences between men and women; namely, the female subjects were less accurate in their responses ("women don't have good heads for figures"),

and they were more likely to change their answers. Figure 22 shows the comparison with respect to the likelihood of change for men and women. At any distance from the median, female subjects were more likely to change than male subjects.

Table 7 shows the relative performance with respect to accuracy of men and women. The entries are in terms of the percentile of the average score for the subgroup. Except for the anomalous case of female scientists (represented by only one individual), the percentile scores for the women are uniformly lower than those for the men.

One possibility is that the differences can be accounted for by some other factor than sex. One candidate is intelligence test scores.* In general, the distribution of CMT scores for women was displaced downward. The average CMT score for males was 105, whereas it was 88 for females.** When men and women are compared at the same CMT level, the differences in accuracy remain. However, the difference in changeability is much less noticeable at the higher levels of CMT scores. There doesn't appear to be a good explanation for this anomaly.

Another possibility is that accuracy and changeability are related—i.e., the females are both less accurate and less sure of themselves. This would suggest that there is in general a relation between accuracy and changeability for both men and women. Figure 23 shows the data analysed from this point of view. The hypothesis is borne out, but the women still show a greater amount of changeability than the men at a given level of accuracy.

---

*The Terman Concept Mastery Test(CMT), Form T, was administered to all subjects in four of the experiments.

**These are raw scores, not IQ scores.

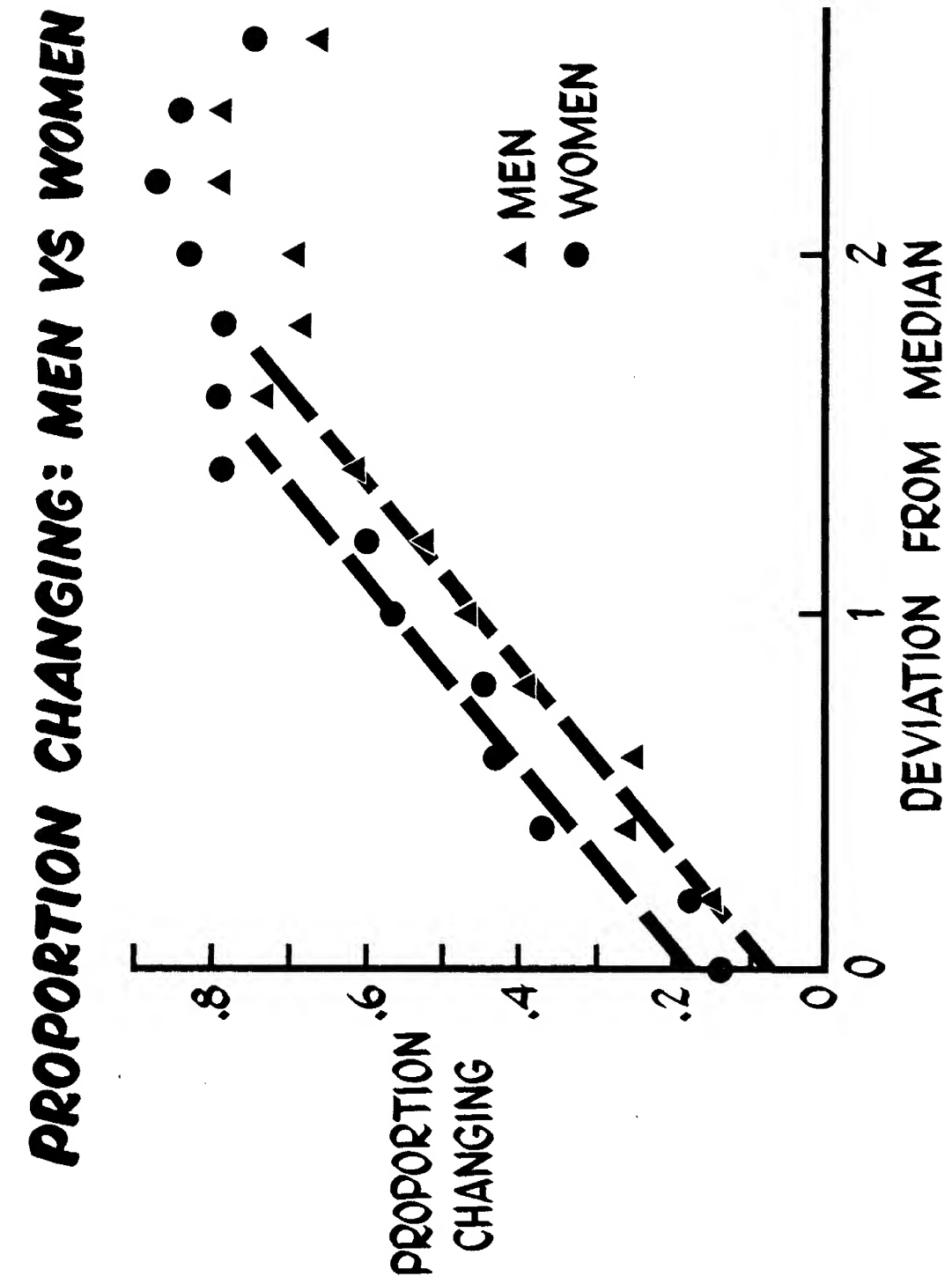
PROPORTION CHANGING: MEN VS WOMEN
PROPORTION CHANGING
.8
.6
.4
.2
0
1
2
DEVIATION FROM MEDIAN
MEN
WOMEN

Fig. 22

Table 7

PERFORMANCE: MAJOR, SEX (Percentile Scores)

MAJOR	MALE	FEMALE
Physical Sciences	44	71*
Biological Sciences	62	41
Psychology	50	36
Economics	62	42
Social Sciences	57	40
Humanities	75	26

*One case.

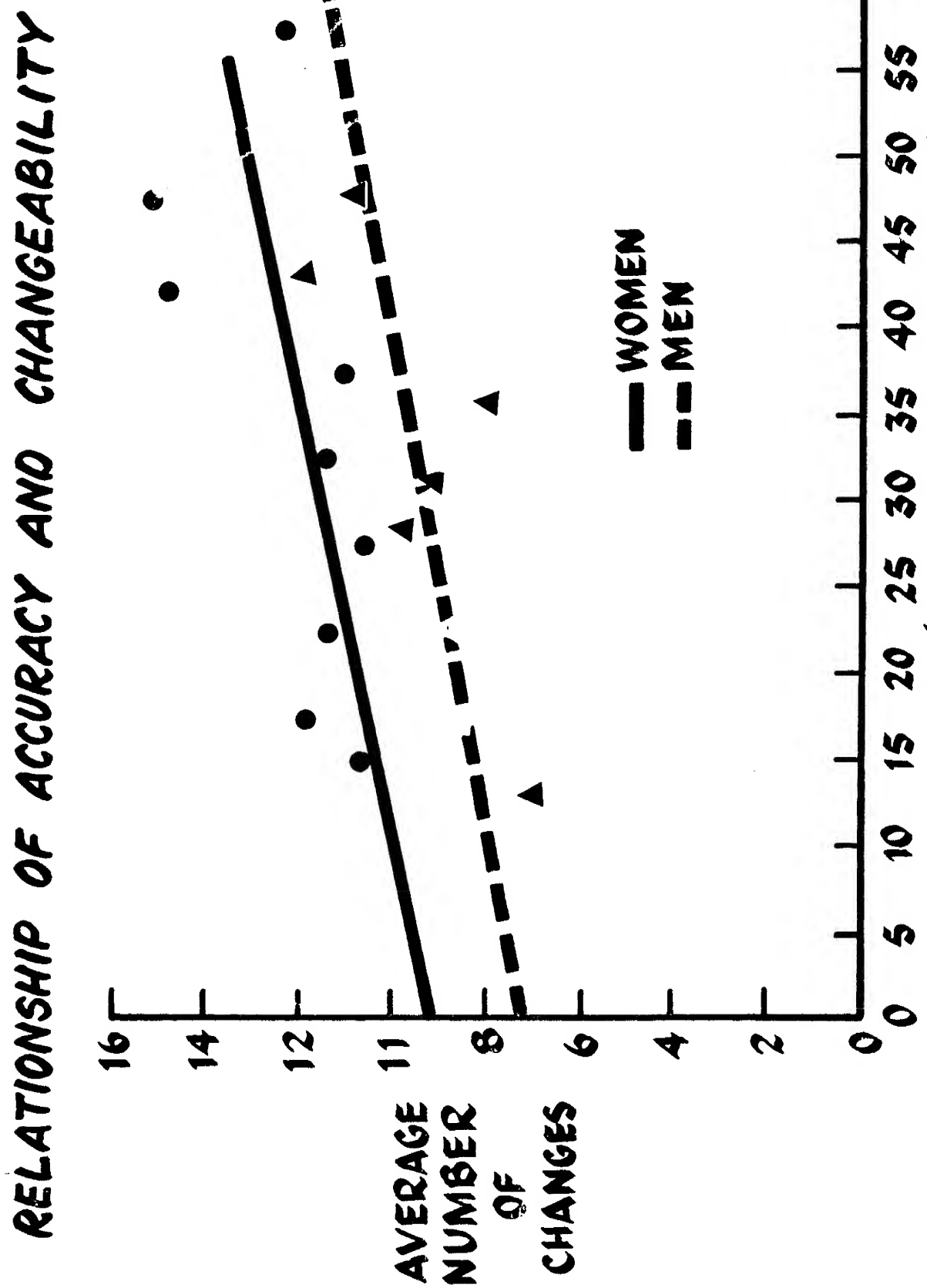
RELATIONSHIP OF ACCURACY AND CHANGEABILITY
AVERAGE NUMBER OF CHANGES
16
14
12
11
8
6
4
2
0
0 5 10 15 20 25 30 35 40 45 50 55 60
ACCURACY (SUM ON 20 QUESTIONS)
WOMEN
MEN

Fig. 23

We conclude that there is an identifiable difference between men and women on both accuracy and changeability. Whether this difference is "cultural" or more basic is, of course, open to conjecture.

9.5. Differences Due to Major. Table 7 also indicates the differences in percentile scores as a function of college major. The results were completely contrary to expectations. It had been expected that students whose major subject was one of the hard sciences would produce more accurate estimates than students from the humanities. In fact, the reverse is the case.

This result adds one more piece of evidence to the presumption that the realm of opinion is different from the realm of knowledge, and that methods which are appropriate for the latter may not be effective in the former.

9.6. Comparison With Simple Iteration. One obvious question is whether the improvement attendant on feedback is simply the result of iteration—rethinking. Two experiments were devoted to investigating this possibility. In the first, twenty-four hours intervened between round one and round two. In the second, a half hour intervened. The first experiment involved a control group that received feedback of first-round medians and quartiles on round two. The second experiment did not involve a control group. In each experiment, a third round with standard feedback was conducted.

The results are displayed in Table 8.

Table 8
EFFECT OF ITERATION WITH AND WITHOUT FEEDBACK

			Round 2 vs Round 1	Round 3 vs Round 2	Round 3 vs Round 1
First Experiment	Without Feedback On Round 2	Better	9	6	9
		Same	2	8	2
		Worse	9	6	9
	With Feedback On Round 2	Better	8	5	10
		Same	7	11	4
		Worse	5	4	6
Second Experiment	Without Feedback On Round 2	Better	4	9	10
		Same	9	6	4
		Worse	7	5	6

The table shows clearly that without feedback there is either no improvement or a degradation. The same groups showed definite improvement with feedback.

9.7. Time. Three experimental sessions were devoted to examining the effect on accuracy of the amount of time allowed to answer. The time intervals used for these tests were 15, 30, 60, 120, and 240 seconds. The number of questions involved was 20, 30, 30, 30, and 10, for the respective time intervals. The results of these tests are plotted in Fig. 24. The point for 240 seconds is omitted because of the small number of questions involved. The plot shows a minimum for the average error in the vicinity of 30 seconds. Performance at the shortest time allowed, 15 seconds, is somewhat poorer than at 1 or 2 minutes.

The most significant feature of the results is the

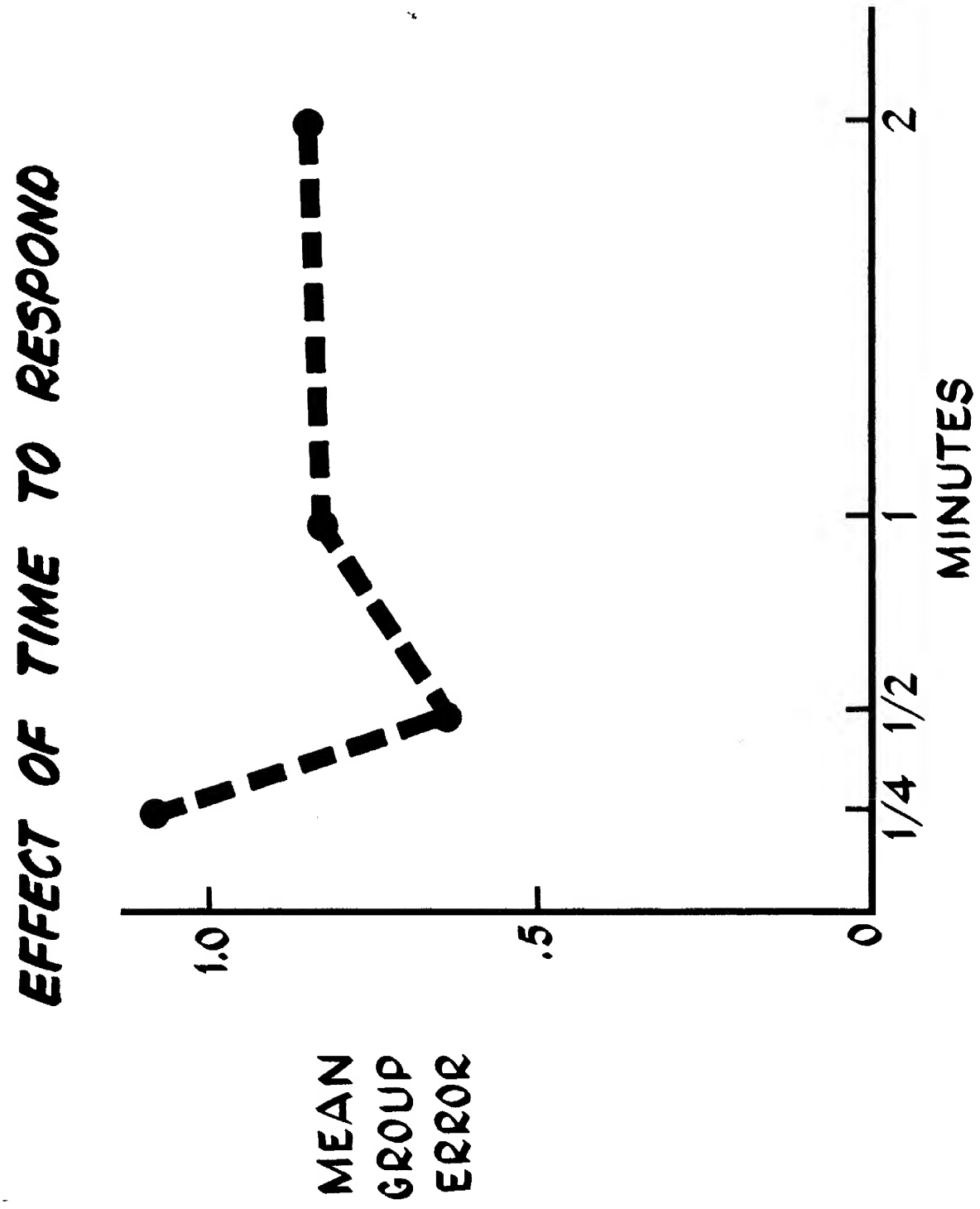
EFFECT OF TIME TO RESPOND
MEAN GROUP ERROR
1.0
.5
0
1/4
1/2
1
2
MINUTES

Fig. 24

occurence of a minimum. Fifteen seconds is barely enough time for the subjects to read a question and write down an answer; so there is no surprise that errors were high for this case. However, even the rather simple almanac-type questions we employed can involve a comparatively complex judgment. A question like "What was the popular vote for Kennedy in the 1960 presidential election in the state of Texas?" involves a number of factors: the population of Texas in 1960, the facts that Texas is a southern state, preponderantly Democratic, but conservative, and predominantly Protestant, that Kennedy was Catholic, and so on. Apparently, there is a fairly sharp limit on the number of factors and the amount of "processing" that can be dealt with profitably. At all events, we seem to have validated the advice frequently given in connection with objective examinations—"Trust your first estimate."

9.8. Self-Evaluation. In several of the experiments, subjects were asked to rate their answers in terms of either their confidence in their responses or their relative competence. Generally, a nominal scale of integers from 1 to 5 was used for these ratings. The basic hypothesis being tested was that a subgroup of more knowledgeable individuals could be selected in terms of their self-rating, and that this subgroup would in general be more accurate than the total group. In every case this hypothesis was not confirmed. In addition, the correlation between accuracy and confidence, or self-rated competence, was extremely variable among the groups, ranging from .65 to .07.

On the other hand, the group reliability for average self-confidence on individual questions was quite high. This was measured, for those cases where there was a control group, by correlating the average self-confidence of one group with the average self-confidence of the other over the set of 20 questions. Reliabilities ranged from .95 to .60

with a mean of .81. In short, the self-confidence ratings appear to be measuring something about the questions fairly well and not just individual differences in self-assurance.

Figure 25 shows the relationship between the group average of self-ratings and mean group error (curve approximated by hand). There is a clear inverse relationship between the group self-rating and group error—in short, the higher the average confidence rating on a question, the smaller the group error. This result has two implications. Although individual self-ratings do not seem to be sufficiently accurate to allow the selection of a more competent subgroup, the average of individual self-ratings (the "group self-rating") appears to be a useful indicator of the accuracy of the group answer. The second implication may be more significant. It was stated in the introduction that one of the major stumbling blocks in dealing with opinion was the lack of a suitable measure for the "solidity"—i.e., the degree of verification—of an assertion in the opinion area. It does not seem unreasonable a priori that group judgments of the degree of verification should be about as accurate and reliable as the group estimates themselves. Figure 25 bears out this presumption.

Table 9 shows the results of combining the dispersion-accuracy relationship and the group rating-accuracy relationship. Dispersion and group rating have a correlation of only .40; thus, there is the possibility that they can operate as separate discriminators with respect to accuracy. Table 9 indicates that this is indeed the case. For a fixed standard deviation, accuracy increases with increasing group rating, and for a fixed group rating, accuracy decreases with increasing standard deviation. The anomalies in the lower right hand boxes may be accounted for by thin statistics; the number of cases is indicated in the small interior boxes.

Table 9 shows that a combination of group self-rating and standard deviation furnishes a relatively sensitive

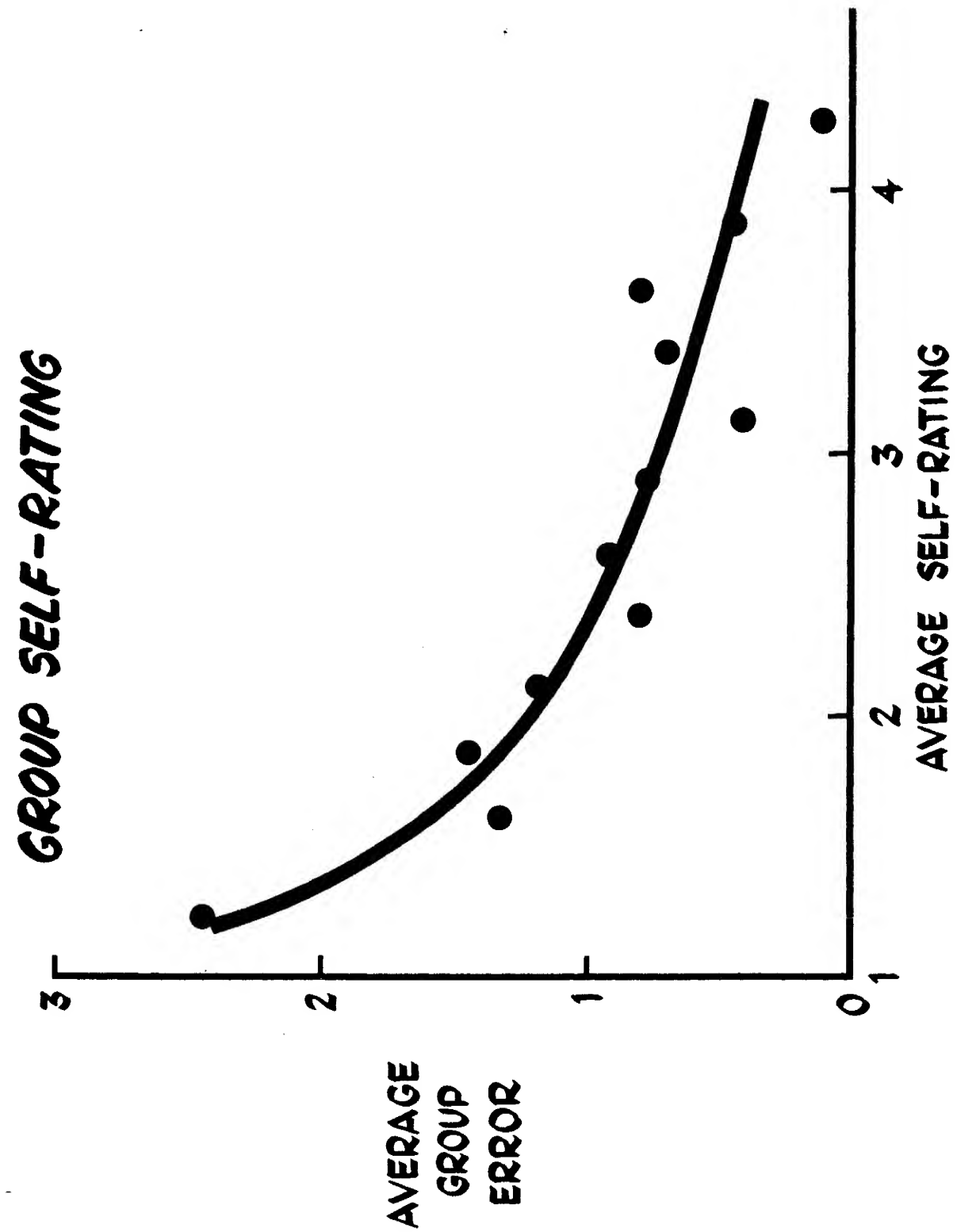
GROUP SELF-RATING
AVERAGE GROUP ERROR
AVERAGE SELF-RATING
0
1
2
3
4

Fig. 25

Table 9

GROUP ERROR AS A FUNCTION OF STANDARD DEVIATION AND GROUP SELF-RATING

Round 1 σ / Group Self Rating	0-.49	.50-.99	1.00-1.49	1.5 up
1-1.99		1.386 [1]	1.114 [11]	1.706 [26]
2-2.49		.787 [4]	.843 [14]	1.106 [36]
2.5-2.99	.655 [1]	.651 [9]	.767 [9]	1.083 [22]
3 up	.139 [15]	.339 [14]	.966 [10]	1.578 [8]

measure of the average accuracy of the group response. The significance of this result can hardly be overemphasized. It opens the possibility that these two parameters can furnish a practical (albeit statistical) measure of the "solidity" of the outputs of a Delphi exercise.

10. DELPHI AND VALUE JUDGMENTS

In policy formulation and decisionmaking, two different kinds of inputs are involved. One is factual judgments, and the other is value judgments. The experimental work on Delphi procedures has dealt exclusively with factual judgments. However, in applications, Delphi procedures have been employed to elicit and process value judgments. A fairly popular form of value judgment is the formulation of the major objectives of an organization and the weighting of these objectives on some scale—e.g., the allocation of 100 points among the objectives.

As far as the workability of Delphi for such value judgments is concerned—in the sense that respondents are willing to furnish lists of objectives, to allocate weights, and to accept a statistical aggregation of weights supplied by a group—the procedures appear to be feasible. But the question of the validity of the procedures is much more obscure when value judgments are involved. The prevailing opinion at the present time appears to be that there is no clear sense in which value judgments can be said to be true or accurate.* Hence, it is of practical importance to ask whether there is any objective way to test Delphi procedures in the value area.

---

*I do not agree with this prevailing opinion. My own opinion is that value judgments are factual statements of an especially complex, vague, and in general much more speculative sort than the usual descriptive inputs to decision situations. However, the demonstration of this point of view is extremely difficult, and not likely to get out of the realm of controversy in the near future. Luckily, as in most intellectual endeavors, controversy about foundations is not incompatible with progress.

The issue is somewhat paradoxical. It is difficult to believe that when a group of corporate policy makers formulate a set of objectives for a major industrial firm, they would accept the judgment that any other set of objectives would be just as good as the set they have produced. In this respect, there is apparently some sense in which they presume that their list is "correct." It would seem that without some such weak presumption, the making of value judgments is rather futile. However, this unpleasant possibility cannot be rejected a priori.

With the weak assumption that there is a "correct" answer that the group is trying to estimate, most of the discussion in Section 2 becomes applicable. If the judgment can be expressed in numerical terms, as for example the weights to be placed on objectives, then, in the absence of ways of distinguishing among a group of respondents with respect to their value-judgment-making ability, the group response is at least as likely to be "correct" as that of half the respondents. In addition, the comments on reliability carry over if the distribution of value numbers is not pathological.

This is a somewhat surprising conclusion, considering the usual feeling that value judgments are nebulous and "unmanageable." The basic assumption, however, is not vacuous. There are three testable consequences of the hypothesis that there is a "correct" judgment: (1) Individual judgments cannot be capricious in the sense that they "could be anything." This is a difficult consequence to test directly. It requires that individual judgments have a reasonable amount of stability. But a simple retest for reliability runs into the problem of memory. It an individual expresses a given judgment at a particular time and is asked the same question some time later, he is very likely to remember his previous answer, thus introducing a spurious reliability. However, the consequence can be

tested indirectly taking into account the group distribution of answers. If the distribution of answers is "reasonable"—e.g., not completely flat, or not U-shaped, etc.—the hypothesis that the responses are not capricious receives some confirmation. (2) The group should exhibit convergence given iteration with feedback. In part, this requirement is set by analogy with factual judgments, and in part by the consideration that, if there is a judgment that the group is trying to approximate, then individual judgments should be influenced in a reasonable way by the additional information furnished by feedback from the group.* (3) Judgments should exhibit a reasonable amount of group reliability—i.e., two highly similar groups should, on the whole, arrive at similar judgments and on iteration should move in the same direction.

All of the above requirements are, of course, statistical; and it is to be expected that they would be violated for some judgments. However, if the requirements are not met for a majority of judgments, it seems reasonable to assume that there is very little substance to the judgments. The three consequences listed above are open to experimental test and will be tested in a series of experiments to be initiated in the near future.

---

*There is some evidence that when individuals are expressing personal value judgments—i.e., essentially saying "this is how I feel about this question," there is very little convergence attendant upon feedback.

11. COMMENTS

The experiments described above were conducted in the hope that they would shed some light on the area of opinion as it enters into decisionmaking. The experimental subject matter, and the subjects, do not precisely match the area of interest. Obviously, any of the results must be interpreted carefully before being applied to substantive exercises involving "experts." For example, with almanac-type questions and student subjects, we found a surprisingly low "optimal" time for answering—somewhat between one-half and one minute. This experiment was conducted in a closed information situation and gives no indication of the optimal "ingestion" time for a new piece of hard information. It also does not indicate the optimal time for more complex kinds of estimation, e.g., estimates of the time of occurrence of significant technological events. About the most one can derive from the experiment with regard to applications to other types of subject matter is the presumption that there will be a point beyond which there will be diminishing, perhaps negative, returns to further time invested in thinking about the estimate; and probably this point of dimishing returns will be lower than is normally supposed.

In general, the most significant parameter is likely to be how much the individual members know about the subject matter. Unfortunately, this is the least controllable variable in the experimental situation. It is also difficult to assess in application. The experiments suggest that it is no great loss to include less knowledgeable individuals, since they are more likely to improve on iteration than the more informed (or at least the more accurate) individuals.

There is a reasonable general "theoretical" justification for thinking that inclusion of less knowledgeable individuals in the group is desirable, with some "ifs" thrown in. Figure 26 is a highly schematic illustration of the

SCHEMATIC DIAGRAM OF RELATIVE KNOWLEDGABILITY

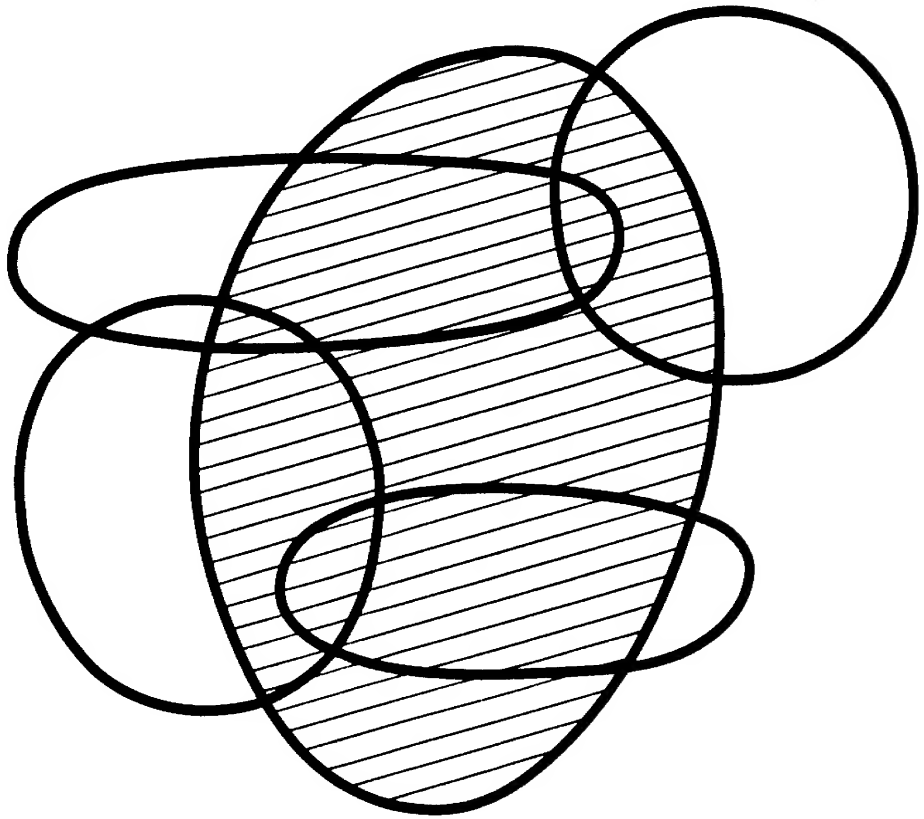

Fig. 26

situation. To make the point clear, assume there is one especially knowledgeable individual (cross-hatched in Fig. 26) and a group of less knowledgeable members. Ordinarily, the knowledgeable one does not have all available information at his command. The response of the more knowledgeable, then, might be better than that of any other single individual but would be worse than what would result if all the information of the group could be pooled. The "if" is important. Our experiments suggest that the residual information of the less knowledgeable does come into play in the iteration step; but we have no way of determining whether anything like the total amount of information gets into the act.

A similar caution needs to be observed with respect to the finding that most forms of feedback beyond the simple statistical report of responses on the previous round are at best ineffective. No experiments were conducted in which hard information was fed back—the auxiliary feedback was in the form of additional opinions. It does not seem unreasonable that rapidly diminishing returns would set in by piling opinion on opinion; but this may not be true for hard information. Here is clearly an important area remaining for further experiment.

REFERENCES

1. Maier, Norman R. F., "Assets and Liabilities in Group Problem Solving: The Need for an Integrative Function," Psychological Review, Vol. 74, No. 4, July 1967, pp. 239-249.

2. Kelly, H. H., and J. W. Thibaut, "Experimental Studies of Group Problem Solving and Process," Gardner Lindzey (ed.), Handbook of Social Psychology, Vol. II Addison-Wesley Publishing Company, Inc., Reading, Mass., 1954.

3. Asch, S. E., "Effects of Group Pressure Upon the Modification and Distortion of Judgments," Eleanor E. Maccoby, etal., (eds.), Readings in Social Psychology, 3rd ed., Holt, Rinehardt and Winston, London, 1958.

4. Girshick, M., A. Kaplan, and A. Skogstad, "The Prediction of Social and Technological Events," Public Opinion Quarterly, Spring 1950, pp. 93-110.

5. Dalkey, N., and O. Helmer, "An Experimental Application of the Delphi Method to the Use of Experts," Management Science, Vol. 9, No. 3, April 1963, pp. 458-467.

6. Gordon, T., and O. Helmer, Report on a Long-Range Forecasting Study, The RAND Corporation, P-2982 (DDC No. AD607777), September 1964.

7. North, Harper Q., "Technology, the Chicken--Corporate Goals, the Egg," Technological Forecasting for Industry and Government, James R. Bright (ed.), Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1968.

8. Campbell, R., "A Methodological Study of the Utilization of Experts in Business Forecasting," unpublished Ph.D. dissertation, UCLA, 1966.

9. Brown, B., S. Cochran, and N. Dalkey Experiments in Group Estimation, The RAND Corporation, RM-5957-PR (in preparation)

10. Helmer, O. Systematic Use of Expert Opinions, The RAND Corporation, P-3721 (DDC No. AD662320), November 1967.